Neural Inverse Rendering of an Indoor Scene From a Single Image

Soumyadip Sengupta[1,2,3,†], Jinwei Gu[1,4,†], Kihwan Kim[1], Guilin Liu[1], David W. Jacobs[2], and Jan Kautz[1]

[1]NVIDIA, [2]University of Maryland, College Park, [3]University of Washington, [4]SenseTime

Abstract

Inverse rendering aims to estimate physical attributes of a scene, e.g., *reflectance, geometry, and lighting, from image(s). Inverse rendering has been studied primarily for single objects or with methods that solve for only one of the scene attributes. We propose the first learning based approach that jointly estimates albedo, normals, and lighting of an indoor scene from a single image. Our key contribution is the Residual Appearance Renderer (RAR), which can be trained to synthesize complex appearance effects (*e.g., *inter-reflection, cast shadows, near-field illumination, and realistic shading), which would be neglected otherwise. This enables us to perform self-supervised learning on real data using a reconstruction loss, based on re-synthesizing the input image from the estimated components. We finetune with real data after pretraining with synthetic data. To this end we use physically-based rendering to create a large-scale synthetic dataset, which is a significant improvement over prior datasets. Experimental results show that our approach outperforms state-of-the-art methods that estimate one or more scene attributes.*

1. Introduction

Inverse rendering aims to estimate physical attributes (*e.g.*, geometry, reflectance, and illumination) of a scene from images. It is one of the core problems in computer vision, with a wide range of applications in gaming, AR/VR, and robotics [12, 14, 40, 43]. In this paper, we propose a holistic, data-driven approach for inverse rendering of an indoor scene from a single image. Inverse rendering has two main challenges. First, it is inherently ill-posed, especially if only a single image is given. Previous approaches [1, 15, 23, 25] that aim to solve this problem from a single image focus only on a single object. Second, inverse rendering of a *scene* is particularly challenging, compared to single objects, due to complex appearance effects (*e.g.*, inter-reflection, cast shadows, near-field illumination, and realistic shading). Some existing works [8, 21, 46, 19] are limited to estimating only one of the scene attributes. We focus on jointly estimating all scene attributes from a single image, which outperforms previous approaches that estimate a single attribute and generalizes better across datasets (see Figure 1 and more in Section 5).

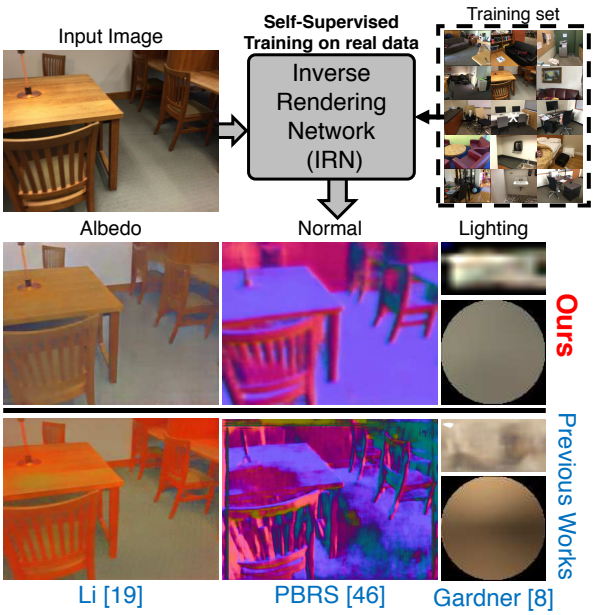

Figure 1: We propose a self-supervised approach for inverse rendering. We jointly decompose an indoor scene image into albedo, surface normal and environment map lighting (top). Our method outperforms state-of-the-art approaches (bottom) that solve for only one of the scene attributes, *i.e.* albedo (Li [19]), normal (PBRS [46]) and lighting (Gardner [8]).

A major challenge in solving this problem is the lack of ground-truth labels for real images. Although we have ground-truth labels for geometry, collected by depth sensors, it is extremely difficult to measure reflectance and lighting at the large scale needed for training a neural network. Networks trained on synthetic images often fail to generalize well on real images. In this paper, we propose two key innovations aimed to tackle the domain gap between synthetic and real images. First, we propose the Residual Appearance Renderer (RAR), which aims to model complex appearance effects by predicting the resid-

†The authors contributed to this work when they were at NVIDIA.

uals that can not be captured with our distant illumination model. This enables us to learn from unlabeled real images using self-supervised photometric reconstruction loss by re-synthesizing the image from its estimated components. Second, we introduce a new synthetic dataset using physically-based rendering with more photorealistic images than previous indoor scene datasets [46, 37]. More realistic synthetic data is useful for pretraining our network to help bridge the domain gap between real and synthetic images.

Residual Appearance Renderer. Our key idea is to learn from unlabeled real data using self-supervised reconstruction loss, which is enabled by our proposed Residual Appearance Renderer (RAR) module. While using a reconstruction loss for domain transfer from synthetic to real has been explored previously [32, 36, 23], their renderer is limited to direct illumination under distant lighting with a single material. For real images of a scene, however, this simple direct illumination renderer cannot synthesize important, complex appearance effects, such as inter-reflections, cast shadows, near-field lighting, and realistic shading. These effects termed *residual appearance* in this paper, can only be simulated with the rendering equation via ray-tracing, which is non-differentiable and is extremely difficult to employ in a learning-based framework. To this end, we propose the Residual Appearance Renderer (RAR) module, which along with a direct illumination renderer can reconstruct the original image from the estimated scene attributes, for self-supervised learning on real images.

Rendering dataset. It is especially challenging for inverse rendering tasks to obtain accurate ground truth labels for real images. Hence we synthesize a large-scale dataset, *CG-PBR*, by applying physically-based rendering to all the 3D indoor scenes from SUNCG [37]. Compared to prior work (PBRS [46]), we significantly improve data quality in the following ways: (1) The rendering of a scene is performed under multiple natural illuminations. (2) We render the same scene twice. Once with all materials set to Lambertian and once with the default material settings to produce image pairs (diffuse, specular). (3) We utilize deep denoising [3], which allows us to render high-quality images from limited samples per pixel. Our dataset consists of 235,893 images with labels for normal, depth, albedo, Phong [16] model parameters and semantic segmentation. Examples are shown in Fig. 4.

Similar to prior works [19, 48], we also make use of sparse labels on real data [2, 27], whenever available, to further improve performance on real images.

To our knowledge, our approach is the first data-driven solution to single-image based inverse rendering of an indoor scene. SIRFS [1], which is a classical optimization based method, seems to be the only prior work with similar goals. Compared with SIRFS, as shown in Sec. 5, our method is more robust and accurate. In addition, we also compare with recent DL-based methods that estimate only one of the scene attributes, such as albedo [19, 20, 28, 48], lighting [8], and normals [46]. Both quantitative and qualitative evaluations show that our approach outperforms these single-attribute methods and generalizes better across datasets, which seems to indicate that the self-supervised joint learning of all these scene attributes is helpful. Our code is available for research purposes here.

2. Related Work

Optimization-based approaches. For inverse rendering from a few images, most traditional optimization-based approaches make strong assumptions about statistical priors on illumination and/or reflectance. A variety of sub-problems have been studied, such as intrinsic image decomposition [39], shape from shading [30, 29], and BRDF estimation [24]. Recently, SIRFS [1] showed it is possible to factorize an image of an object or a scene into surface normals, albedo, and spherical harmonics lighting. In [33] the authors use CNNs to predict the initial depth and then solve inverse rendering with an optimization. From an RGBD video, Zhang *et al*. [44] proposed an optimization method to obtain reflectance and illumination of an indoor room. These optimization-based methods, although physically grounded, often do not generalize well to real images where those statistical priors are no longer valid.

Learning-based approaches. With recent advances in deep learning, researchers have proposed to learn data-driven priors to solve some of these inverse problems with CNNs, many of which have achieved promising results. For example, it is shown that depth and normals may be estimated from a single image [6, 7, 49] or multiple images [38]. Parametric BRDF may be estimated either from an RGBD sequence of an object [25, 15] or for planar surfaces [21]. Lighting may also be estimated from images, either as an environment map [8, 10], or spherical harmonics [47] or point lights [45]. Recent works [36, 32] also performed inverse rendering on faces. Some recent works also jointly learn some of the intrinsic components of an object, like reflectance and illumination [9, 41], reflectance and shape [22], and normal, BRDF, and distant lighting [34, 23]. Nevertheless, these efforts are mainly limited to objects rather than scenes, and do not model the aforementioned residual appearance effects such as inter-reflection, near-field lighting, and cast shadows present in real images.

Differentiable Renderer. A few recent works from the graphics community proposed differentiable Monte Carlo renderers [18, 4] for optimizing rendering parameters (*e.g*., camera poses, scattering parameters) for synthetic 3D scenes. Neural mesh renderer [13] addressed the problem of differentiable visibility and rasterization. Our proposed RAR is in the same spirit, but its goal is to synthesize the

complex appearance effects for inverse rendering on *real images*, which is significantly more challenging.

Datasets for inverse rendering. High-quality synthetic data is essential for learning-based inverse rendering. SUNCG [37] created a large-scale 3D indoor scene dataset. The images of the SUNCG dataset are not photo-realistic as they are rendered with OpenGL using diffuse materials and point source lighting. An extension of this dataset, PBRS [46], uses physically based rendering to generate photo-realistic images. However, due to the computational bottleneck in ray-tracing, the rendered images are quite noisy and limited to one lighting condition. There also exist a few real-world datasets with partial labels on geometry, reflectance, or lighting. NYUv2 [27] provides surface normals from indoor scenes. Relative reflectance judgments from humans are provided in the IIW dataset [2] which are used in many intrinsic image decomposition methods. In contrast to these works, we created a large-scale synthetic dataset with significant image quality improvement.

Intrinsic image decomposition. Intrinsic image decomposition is a sub-problem of inverse rendering, where a single image is decomposed into albedo and shading. Recent methods learn intrinsic image decomposition from labeled synthetic data [17, 26, 34] and from unlabeled [20] or partially labeled real data [48, 19, 28, 2]. Intrinsic image decomposition methods do not explicitly recover geometry or illumination but rather combine them together as shading. In contrast, our goal is to perform a complete inverse rendering which has a wider range of applications in AR/VR.

3. Our Approach

We present a deep learning-based approach for inverse rendering from a single image, which is shown in Fig. 2. Given an input image I, our proposed neural Inverse Rendering Network (IRN), denoted as $h_d(\cdot;\Theta_d)$, estimates surface normal N, albedo A, and environment map L:

$$\text{IRN}: \quad h_d(I;\Theta_d) \rightarrow \left\{\hat{A}, \hat{N}, \hat{L}\right\}. \tag{1}$$

Using our synthetic data CG-PBR, we can simply train IRN ($h_d(\cdot;\Theta_d)$ with supervised learning – with only one caveat, *i.e.* we need to approximate the "ground truth" environment maps (using a separate network $h_e(\cdot;\Theta_e) \rightarrow \hat{L}^*$; see Sec. 3.1 for details). To generalize on real images, we use a self-supervised reconstruction loss. Specifically, as shown in Fig. 2, we use two renderers to re-synthesize the input image from the estimations. The direct renderer $f_d(\cdot)$ is a simple closed-form shading function with no learnable parameters, which synthesizes the direct illumination part $\hat{I}_d$ of the the raytraced image. The Residual Appearance Renderer (RAR), denoted by $f_r(\cdot;\Theta_r)$, is a trainable network module, which learns to synthesize the complex appearance effects $\hat{I}_r$:

$$\text{Direct Renderer}: \quad f_d(\hat{A}, \hat{N}, \hat{L}) \rightarrow \hat{I}_d \tag{2}$$

$$\text{RAR}: \quad f_r(I, \hat{A}, \hat{N}; \Theta_r) \rightarrow \hat{I}_r. \tag{3}$$

The self-supervised reconstruction loss is thus defined as $||I - (\hat{I}_d + \hat{I}_r)||_1$. We explain the details of the direct renderer and the RAR in Sec. 3.2.

3.1. Training on Synthetic Data

We first train IRN on our synthetic dataset CG-PBR with supervised learning. As shown in Fig. 2, IRN has a structure similar to [32], which consists of a convolutional encoder, followed by nine residual blocks and a convolutional decoder for estimating albedo and normals. We condition the lighting estimation block on the image, normals and albedo features. We use ground truth albedos A^* and normals N^* from CG-PBR for supervised learning.

The ground truth environmental lighting L^* is challenging to obtain, as it is the approximation of the actual surface light field. We use environment maps as the exterior lighting for rendering CG-PBR, but these environment maps cannot be directly set as L^*, because the virtual cameras are placed *inside* each of the indoor scenes. Due to occlusions, only a small fraction of the exterior lighting (*e.g.*, through windows and open doors) is directly visible. The surface light field of each scene is mainly due to global illumination and some interior lighting. One could approximate L^* by minimizing the difference between the raytraced image I and the output I_d of the direct renderer $f_d(\cdot)$ with ground truth albedo A^* and normal N^*. However, we found this approximation to be inaccurate, since $f_d(\cdot)$ cannot model the residual appearance present in the raytraced image I.

We thus resort to a learning-based method to approximate the ground truth lighting L^*. Specifically, we train a residual block based network, $h_e(\cdot;\Theta_e)$, to predict $\hat{L}^*$ from the input image I, normals N^* and albedo A^*. We first train $h_e(\cdot;\Theta'_e)$ with the images synthesized by the direct renderer $f_d(\cdot)$ with ground truth normals, albedo and indoor lighting, $I_d = f_d(A^*, N^*, L)$, where L is randomly sampled from a set of real *indoor* environment maps. Here the network learns a prior over the distribution of indoor lighting, *i.e.*, $h(I_d;\Theta'_e) \rightarrow L$. Next, we fine-tune this network $h_e(\cdot;\Theta'_e)$ on the raytraced images I, by minimizing the reconstruction loss: $\|I - f_d(A^*, N^*, \hat{L}^*)\|$. Thus we obtain the approximated ground truth of the environmental lighting $\hat{L}^* = h_e(I;\Theta_e)$ which can best reconstruct the raytraced image I modelled by the direct render.

Finally, with all the ground truth components ready, the supervised loss for training IRN is

$$\begin{aligned} L_s = {} & \lambda_1||\hat{N} - N^*||_1 + \lambda_2||\hat{A} - A^*||_1 \\ & + \lambda_3||f_d(A^*, N^*, \hat{L}) - f_d(A^*, N^*, \hat{L}^*)||_1. \end{aligned} \tag{4}$$

where $\lambda_1 = 1$, $\lambda_2 = 1$, and $\lambda_3 = 0.5$.

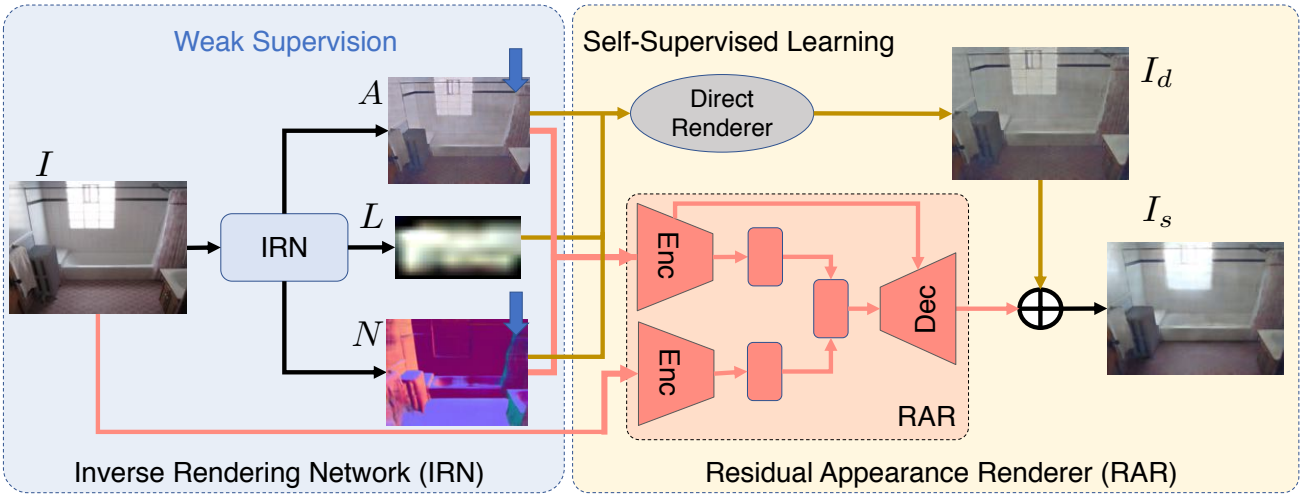

Figure 2: **Overview of our approach.** Our Inverse Rendering Network (IRN) predicts albedo, normals and illumination map. We train on unlabeled real images using self-supervised reconstruction loss. Reconstruction loss consists of a closed-form Direct Renderer with no learnable parameters and the proposed Residual Appearance Renderer (RAR), which learns to predict complex appearance effects.

3.2. RAR: Self-supervised Training on Real Images

Learning from synthetic data alone is not sufficient to perform well on real images. Although CG-PBR was created with physically-based rendering, the variation of objects, materials, and illumination is still limited compared to those in real images. Since obtaining ground truth labels for inverse rendering is almost impossible for real images (especially for reflectance and illumination), we use two key ideas for domain transfer from synthetic to real: (1) self-supervised reconstruction loss and (2) weak supervision from sparse labels.

Previous works on faces [32, 36] and objects [23] have shown success in using a self-supervised reconstruction loss for learning from unlabeled real images. As mentioned earlier, the reconstruction in these prior works is limited to the direct renderer $f_d(\cdot)$, which is a simple closed-form shading function (under distant lighting) with no learnable parameters. In this paper, we implement $f_d(\cdot)$ simply as

$$\hat{I}_d = f_d(\hat{A}, \hat{N}, \hat{L}) = \hat{A} \sum_i \max(0, \hat{N} \cdot \hat{L}_i), \qquad (5)$$

where $\hat{L}_i$ corresponds to the pixels on the environment map $\hat{L}$. While using $f_d(\cdot)$ to compute the reconstruction loss may work well for faces [32] or small objects with homogeneous material [23], we found that it fails for inverse rendering of a scene. In order to synthesize the aforementioned residual appearances (*e.g.*, inter-reflection, cast shadows, near-field lighting), we propose to use the differentiable Residual Appearance Renderer (RAR), $f_r(\cdot; \Theta_r)$, which learns to predict a residual image $\hat{I}_r$. The self-supervised reconstruction loss is thus defined as $L_u = ||I - (\hat{I}_d + \hat{I}_r)||_1$.

Our goal is to train RAR to capture only the residual appearances but *not* to correct the artifacts of the direct rendered image due to faulty normals, albedo, and lighting estimation of the IRN. To achieve this goal, we train RAR *only* on synthetic data with ground-truth normals and albedo, and fix it for training on real data, so that it only learns to correctly predict the residual appearances when the direct renderer reconstruction is accurate. We need realistic synthetic images, which capture complex appearance effects, *e.g.*, cast shadows, inter-reflections etc. for training RAR. Our CG-PBR provides the necessary photo-realism that previous indoor scene datasets [46, 37] lack.

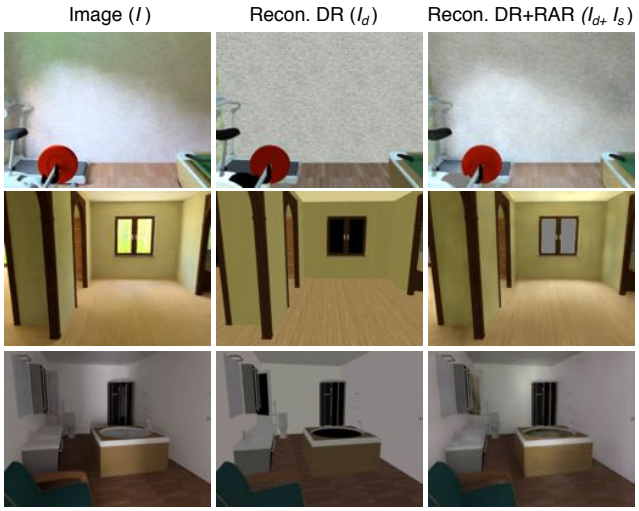

Figure 3: RAR $f_r(\cdot)$ learns to predict complex appearance effects (*e.g.* near-field lighting, cast shadows, inter-reflections) which cannot be modeled by a direct renderer (DR) $f_d(\cdot)$.

As shown in Fig. 2, RAR consists of a U-Net [31], with normals and albedo as its input, and latent image features ($D = 300$ dimension) learned by a convolutional encoder ('Enc'). We combine the image features at the end of the U-Net encoder and process them with the U-Net decoder

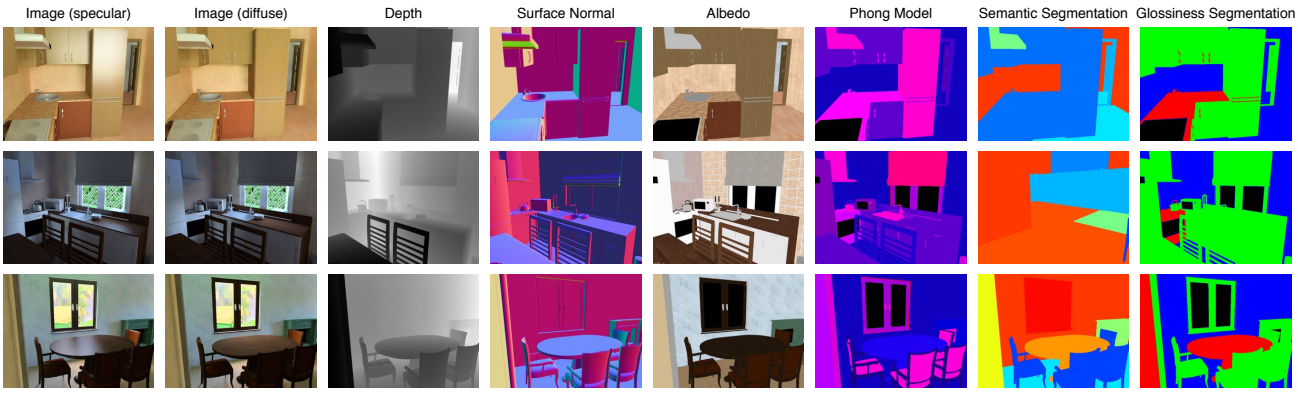

Figure 4: **Our CG-PBR Dataset** consists of 235,893 images of a scene assuming specular and diffuse reflectance along with ground truth depth, surface normals, albedo, Phong model parameters, semantic segmentation and glossiness segmentation.

to produce the residual image. RAR, along with the direct renderer $f_d(\cdot)$, acts like an auto-encoder in principle. RAR learns to encode complex appearance features from the original image into a latent subspace ($D = 300$ dimension). The bottleneck of the auto-encoder architecture present in RAR forces it to focus only on the complex appearance features and not in the whole image. So RAR learns to encode the non-direct part of the image to avoid paying a penalty in the reconstruction loss and in principle is simpler than a differentiable renderer.

As shown in Fig. 3, RAR indeed learns to synthesize complex residual appearance effects present in the original input image. In Sec. 6, we provide quantitative and qualitative ablation studies to show why RAR is crucial in improving albedo and normal estimation. The goal of RAR in this project is to reconstruct an image from its components along with the direct renderer to facilitate self-supervised learning on real images. This helps to significantly improve albedo and normal estimation over state-of-the-art approaches. Our goal is *not* to develop a differentiable renderer for realistic illumination during object insertion.

Similar to prior work [48, 19], we use sparse labels over reflectance as weak supervision during training on real images. Specifically, we use pair-wise relative reflectance judgments from the Intrinsic Image in the Wild (IIW) dataset [2] as a form of supervision over albedo. We also use supervision over surface normals from NYUv2 dataset [27]. As shown in Sec. 6, using such weak supervision can substantially improve performance on real images.

3.3. Training Procedure

We summarize the different stages of training from synthetic to real data.

Estimate GT indoor lighting: (a) First train $h_e(\cdot;\Theta'_e)$ on images rendered by the direct renderer $f_d(\cdot)$. (b) Fine-tune $h_e(\cdot;\Theta_e)$ on raytraced synthetic images to estimate GT indoor environment map $\hat{L}^*$.

Train on synthetic images: (a) Train IRN with supervised L1 loss on albedo, normal and lighting. (b) Train RAR on synthetic data with L1 image reconstruction loss.

Train on real images: Fine-tune IRN on real data with (1) the pseudo-supervision over albedo, normal and lighting (to handle ambiguity of decomposition as proposed in [32]), (2) the self-supervised reconstruction loss L_u with RAR, and (3) the weak supervision over the albedo (*i.e.*, pair-wise relative reflectance judgment) or normals.

4. The CG-PBR Dataset

High-quality synthetic datasets are essential for learning-based inverse rendering. The SUNCG dataset [37] contains 45,622 indoor scenes with 2,644 unique objects, but their images are rendered with OpenGL under fixed point light sources. The PBRS dataset [46] extends the SUNCG dataset by using physically based rendering with Mitsuba [11]. Yet, due to a limited computational budget, many rendered images in PBRS are quite noisy. Moreover, the images in PBRS are rendered with only diffuse materials and a single outdoor environment map, which also significantly limits the photo-realism of the rendered images. High-quality photo-realistic images are necessary for training RAR to capture residual appearances.

In this paper, we use a new dataset named CG-PBR, which improves data quality in the following ways: (1) The rendering is performed under multiple outdoor environment maps. (2) We render the same scene twice, once with all materials set to Lambertian and once with the default material settings. This offers (diffuse, specular) image pairs which can be useful to the community for learning to remove highlights and many other potential applications. (3) We utilize deep denoising [3], which allows us to raytrace high-quality images from limited samples per pixel. Our dataset consists of 235,893 images with labels related to normal, depth, albedo, Phong [16] model parameters, semantic and glossiness segmentation. Examples are shown in Fig. 4. A comparison with the SUNCG and PBRS datasets is shown in Fig. 5.

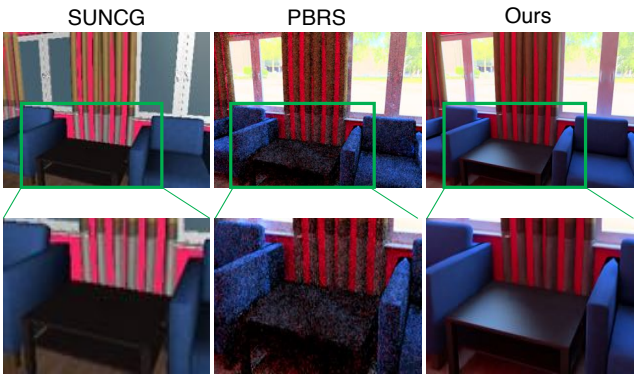

Figure 5: **Comparison with PBRS [46] and SUNCG [37].** Our dataset introduces more photo-realistic and less noisy images with specular highlights under multiple lighting conditions.

5. Experimental Results

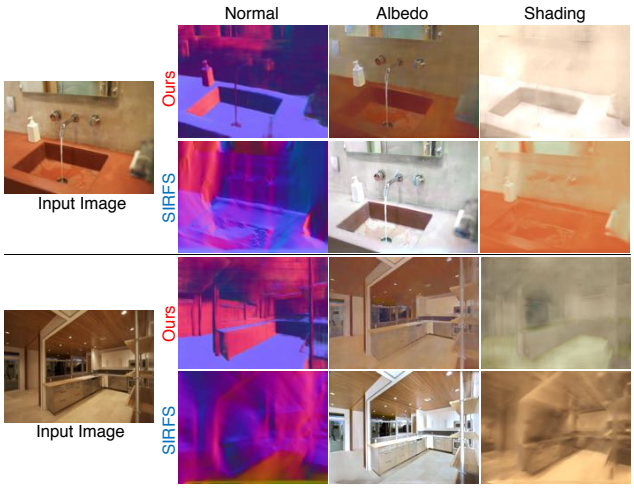

Figure 6: **Comparison with SIRFS [1].** Using deep CNNs our method performs better disambiguation of reflectance from shading and predicts better surface normals.

Comparison with SIRFS. SIRFS [1] is an optimization-based method for inverse rendering, which estimates surface normals, albedo and spherical harmonics lighting from a single image. Although it is primarily developed for objects, we applied it to scene and it works reasonably well. We compare with SIRFS on the test data from the IIW dataset [2]. As shown in Fig. 6, our method produces more accurate normals and better disambiguation of reflectance from shading. This is expected, as we are using deep CNNs, which are known to better learn and utilize statistical priors present in the data than traditional optimization techniques.

Comparison with intrinsic image decomposition algorithms. Intrinsic image decomposition aims to decompose an image into albedo and shading, which is a sub-problem in inverse rendering. Several recent works [2, 48, 28, 19] showed promising results with deep learning. While our goal is to solve the complete inverse rendering problem, we still compare albedo prediction with these latest intrinsic image decomposition methods. We evaluate the WHDR (Weighted Human Disagreement Rate) metric [2] on the test set of the IIW dataset [2] and report the result in Table 1. As shown, we outperform these algorithms that train on the original IIW dataset [2]. Since our goal is not intrinsic image decomposition, we do not train on additional intrinsic image specific datasets and avoid any post-processing as done in Li *et al*. [19].

Figure 7: **Comparison with CGI (Li *et. al.* [19]).** In comparison with CGI [19], our method performs better disambiguation of reflectance from shading and preserves the texture in the albedo.

Table 1: **Intrinsic image decomposition on the IIW test set [2]**

Algorithm	Training set	WHDR
Bell *et. al.* [2]	-	20.6%
Li *et. al.* [20]	-	20.3%
Zhou *et. al.* [48]	IIW	19.9%
Nestmeyer *et. al.* [28]	IIW	19.5%
Li *et. al.* [19]	IIW	17.5%
Ours	IIW	**16.7%**

Table 2: **Albedo estimation on synthetic data.** (RMSE;MAD)

Algorithm	CG-PBR	CGI [19]
Li *et. al.* [19]	0.2873; 0.2427	0.2685; 0.2220
Ours	**0.1567; 0.1300**	**0.2126; 0.1875**

We also present a qualitative comparison of the inferred albedo with Li *et al*. [19] in Figure 7. As shown, our method preserves more detailed texture and has fewer artifacts in the predicted albedo, compared to Li *et al*. [19]. To further illustrate the effectiveness of our method over Li *et al*. [19], we also evaluate albedo estimation error (RMSE and MAD) on the synthetic CG-PBR and CGI datasets [19] in Table

2. While our method uses CG-PBR as synthetic data for training, Li *et al*. [19] uses CGI. Yet we outperform Li *et al*. [19] on both datasets. Thus our method significantly outperforms all prior intrinsic image decomposition algorithms for albedo estimation.

Evaluation of normal estimation. We also compare with PBRS [46] which predicts only surface normals from an image. Both PBRS and 'Ours' are trained on synthetic data and NYUv2 [27] (real data), and then tested on NYUv2, 7-scenes [35], Scannet [5] and synthetic CG-PBR datasets. In Table 3 we evaluate median angular error over the estimated albedo obtained by PBRS and 'Ours'. Our method significantly outperforms PBRS on Scannet and CG-PBR, while slightly improving on 7-Scenes and doing slightly worse on NYUv2. This suggest that while PBRS overfits on NYUv2, our method shows significantly better generalization across datasets. This is because we are jointly reasoning about all components of the scene. Qualitative comparisons of normals predicted by our method and that of PBRS on Scannet dataset are shown in Fig. 8.

Table 3: **Median angular errors for surface normals.** (trained on synthetic and NYUv2)

	NYUv2	7-scenes	Scannet	CG-PBR
PBRS [46]	**15.33°**	25.65°	30.39°	27.84°
Ours	16.92°	**24.54°**	**21.10°**	**18.67°**

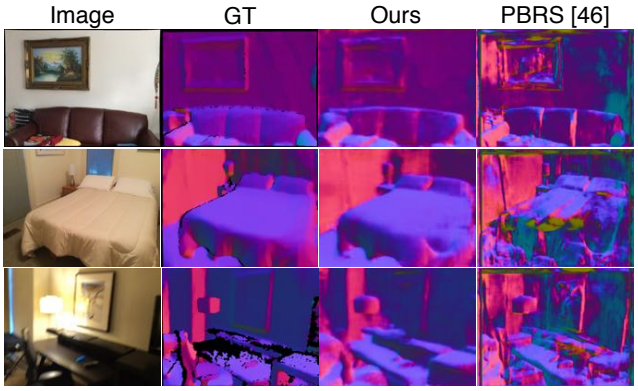

Figure 8: **Comparison with PBRS [46] on Scannet [5].**

Evaluation of lighting estimation. We estimate an environment map of low spatial resolution from an image. We compare our estimated environment map with Gardner *et al*. [8], the only existing method which also estimates a HDR environment map from a single indoor image using CNNs. We present a quantitative evaluation of the estimated environment map on synthetic data in Table 4. For 'Env. map error' we present the Mean Absolute Deviation (MAD) error w.r.t. 'GT' (obtained by using $h_e(\cdot;\Theta_e)$, see Sec. 3.1) weighted by solid angle, following [42]. We also compute MAD 'Image recon. error' by using direct renderer $f_d(\cdot)$ ('GT' has a non-zero image reconstruction error, as it only captures distant-direct illumination.). For evaluating on real data, we collect 20 images of 4 scenes with varying illumination conditions with (also without) a diffuse ball inside the image, which serves as GT, as shown in Fig. 9. We estimate the environment map from the image taken without the diffuse ball using our method and that of Gardner *et al*. [8]. Then we render the diffuse ball with the estimated environment map and evaluate RMSE and MAD reconstruction error with the GT ball in Table 5. Our method significantly outperforms Gardner *et al*. [8] in estimating an environment map from a single image.

Table 4: **Environment map estimation – synth. data.** (MAD)

	Gardner [8]	Ours	GT
Env. map error	0.3162	**0.1639**	-
Image recon. error	0.1640	**0.1158**	0.0949

Table 5: **Environment map estimation – real data.**

	RMSE	MAD
Gardner [8]	0.3109	0.2554
Ours	**0.2105**	**0.1772**

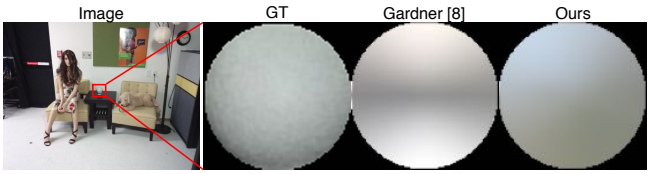

Figure 9: **Comparison with Gardner *et al*. [8].** We collect 20 images of scenes with a diffuse ball, which is cropped as 'GT'. Rendered diffuse ball with environment estimated by 'Ours' outperforms Gardner *et al*. [8] significantly.

6. Ablation Study

Role of RAR in self-supervised training. The goal of RAR is to enable self-supervision on real images by capturing complex appearance effects that cannot be modeled by a direct renderer. RAR, along with the direct renderer, can reconstruct the image from its components, so that it can be used to train with a reconstruction loss. To show the effectiveness of RAR, we train IRN with (a) pseudo-supervision over albedo, normal and lighting (to handle ambiguity of decomposition as proposed in [32]), and (b) weak supervision over normals or albedo, whenever available. Specifically, we train IRN: (i) on IIW with weak-supervision over albedo, with and without RAR; (ii) on NYUv2 with weak-supervision over normals, with and without RAR. Models trained on IIW with supervision over albedo are expected to produce better albedo estimates; models trained on NYUv2 with supervision over normals are expected to produce better normal estimates. We test these models on the CG-PBR and IIW to evaluate albedo and on NYUv2 and Scannet to

evaluate normals. We report MAD error for the CG-PBR, the WHDR measure for the real IIW dataset and median angular error for the NYUv2 and Scannet in Table 6. Overall, RAR significantly improves the albedo and normal estimates across different datasets.

Table 6: **Role of RAR.** We evaluate albedo and normal estimation error by training IRN with ('Ours') and without ('w/o RAR') RAR using weak-supervisions over albedo (IIW) and normal (NYUv2).

	Albedo		Normal	
	CG-PBR	IIW	NYUv2	Scannet
IIW - Ours	**0.130**	**16.7%**	**23.2°**	**28.4°**
w/o RAR	0.265	21.6%	60.2°	91.5°
NYUv2 - Ours	**0.134**	**38.5%**	**16.9°**	**21.1°**
w/o RAR	0.191	40.6%	21.1°	87.1°

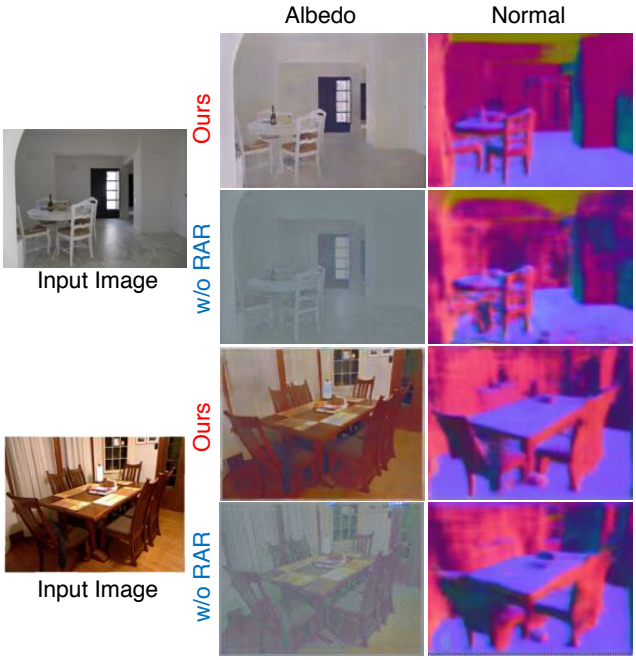

Figure 10: **Role of RAR in self-supervised training.** We train IRN with ('Ours') and without ('w/o RAR') RAR on real data with weak-supervision over normals and albedo. Using RAR significantly improves the quality of estimated albedo and normal.

We further illustrate the importance of RAR with qualitative evaluations in Figure 10. We train IRN with ('Ours') and without RAR ('w/o RAR'), with weak supervision over either albedo (WHDR loss on IIW for predicting albedo) or normal (L1 loss over normals in NYUv2 for predicting normals). Networks trained with supervision over one component, *e.g.* normals, are always expected to produce better estimates of that component (normals) than the other (albedo). The normals are significantly improved when we use RAR. As for the albedo, using relative reflectance judgments without RAR produces very low contrast albedo. In the absence of RAR, the reconstruction loss used for self-supervised training cannot capture complex appearance effects, and thus it produces worse estimates of scene attributes.

Role of weak supervision. To evaluate the influence of weak supervision on inverse rendering, we train IRN with and without weak supervision over albedo (on IIW) and normals (NYUv2), respectively, as shown in Table 7. Weak supervision significantly reduces median angular error on the NYUv2 dataset and the WHDR metric on the IIW dataset. It also makes albedo prediction more consistent across large objects like walls, floors, and ceilings as shown in Fig. 11.

Table 7: **Role of weak supervision.** We train IRN with ('Ours') and without ('w/o RAR') RAR using weak-supervisions over albedo (IIW) and normal respectively (NYUv2).

	Albedo (IIW)	Normal (NYUv2)
Ours	**16.7%**	**16.9°**
w/o weak sup.	32.7%	23.3°

Figure 11: **Role of weak supervision.** We predict more consistent albedo across large objects like walls, floors and ceilings using pair-wise relative reflectance judgments from the IIW dataset [2].

7. Conclusion

We present a learning-based approach for inverse rendering of an indoor scene from a single RGB image. Experimental results show our method outperforms prior works for estimating albedo, normal and lighting, which shows the effectiveness of joint learning. We propose a novel Residual Appearance Renderer (RAR) that can synthesize complex appearance effects such as inter-reflection, cast shadows, near-field illumination, and realistic shading. We show that this renderer is important for employing the self-supervised reconstruction loss to learn inverse rendering on real images. Although the goal of RAR in this paper is to enable self-supervision and improve inverse rendering estimations, we believe it is a good starting point for developing neural renderers that can handle object insertion.

Acknowledgment This research is supported by the National Science Foundation under grant no. IIS-1526234.

References

[1] Jonathan T Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2015. 1, 2, 6

[2] Sean Bell, Kavita Bala, and Noah Snavely. Intrinsic images in the wild. *ACM Transactions on Graphics (TOG)*, 33(4):159, 2014. 2, 3, 5, 6, 8

[3] Chakravarty R Alla Chaitanya, Anton S Kaplanyan, Christoph Schied, Marco Salvi, Aaron Lefohn, Derek Nowrouzezahrai, and Timo Aila. Interactive reconstruction of monte carlo image sequences using a recurrent denoising autoencoder. *ACM Transactions on Graphics (TOG)*, 36(4):98, 2017. 2, 5

[4] Chengqian Che, Fujun Luan, Shuang Zhao, Kavita Bala, and Ioannis Gkioulekas. Inverse transport networks. *arXiv preprint arXiv:1809.10820*, 2018. 2

[5] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5828–5839, 2017. 7

[6] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *International Conference on Computer Vision (ICCV)*, pages 2650–2658, 2015. 2

[7] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[8] Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. Learning to predict indoor illumination from a single image. *ACM Transactions on Graphics (TOG)*, 36(6):176, 2017. 1, 2, 7

[9] Stamatios Georgoulis, Konstantinos Rematas, Tobias Ritschel, Mario Fritz, Luc Van Gool, and Tinne Tuytelaars. DeLight-Net: Decomposing reflectance maps into specular materials and natural illumination. *arXiv preprint arXiv:1603.08240*, 2016. 2

[10] Yannick Hold-Geoffroy, Kalyan Sunkavalli, Sunil Hadap, Emiliano Gambaretto, and Jean-François Lalonde. Deep outdoor illumination estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, 2017. 2

[11] Wenzel Jakob. Mitsuba renderer, 2010. http://www.mitsuba-renderer.org. 5

[12] Kevin Karsch, Varsha Hedau, David Forsyth, and Derek Hoiem. Rendering synthetic objects into legacy photographs. In *ACM Transactions on Graphics (SIGGRAPH Asia)*, pages 157:1–157:12, 2011. 1

[13] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3D mesh renderer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3907–3916, 2018. 2

[14] Erum Arif Khan, Erik Reinhard, Roland W Fleming, and Heinrich H Bülthoff. Image-based material editing. *ACM Transactions on Graphics (TOG)*, 25(3):654–663, 2006. 1

[15] Kihwan Kim, Jinwei Gu, Stephen Tyree, Pavlo Molchanov, Matthias Nießner, and Jan Kautz. A lightweight approach for on-the-fly reflectance estimation. In *International Conference on Computer Vision (ICCV)*, pages 20–28, 2017. 1, 2

[16] Eric P Lafortune and Yves D Willems. Using the modified Phong reflectance model for physically based rendering. 1994. 2, 5

[17] Louis Lettry, Kenneth Vanhoey, and Luc Van Gool. DARN: a deep adversarial residual network for intrinsic image decomposition. In *IEEE Workshop on Applications of Computer Vision (WACV)*, 2018. 3

[18] Tzu-Mao Li, Miika Aittala, Frédo Durand, and Jaakko Lehtinen. Differentiable monte carlo ray tracing through edge sampling. 37(6):222:1–222:11, 2018. 2

[19] Zhengqi Li and Noah Snavely. CGIntrinsics: Better intrinsic image decomposition through physically-based rendering. *European Conference on Computer Vision (ECCV)*, 2018. 1, 2, 3, 5, 6, 7

[20] Zhengqi Li and Noah Snavely. Learning intrinsic image decomposition from watching the world. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 3, 6

[21] Zhengqin Li, Kalyan Sunkavalli, and Manmohan Chandraker. Materials for masses: SVBRDF acquisition with a single mobile phone image. In *European Conference on Computer Vision (ECCV)*, 2018. 1, 2

[22] Zhengqin Li, Zexiang Xu, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Learning to reconstruct shape and spatially-varying reflectance with a single image. In *ACM Transactions on Graphics (SIGGRAPH Asia)*, 2018. 2

[23] Guilin Liu, Duygu Ceylan, Ersin Yumer, Jimei Yang, and Jyh-Ming Lien. Material editing using a physically based rendering network. In *International Conference on Computer Vision (ICCV)*, 2017. 1, 2, 4

[24] Stephen Lombardi and Ko Nishino. Reflectance and illumination recovery in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 38(1):129–141, 2016. 2

[25] Abhimitra Meka, Maxim Maximov, Michael Zollhoefer, Avishek Chatterjee, Hans-Peter Seidel, Christian Richardt, and Christian Theobalt. LIME: Live intrinsic material estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 2

[26] Takuya Narihira, Michael Maire, and Stella X Yu. Direct intrinsics: Learning albedo-shading decomposition by convolutional regression. In *International Conference on Computer Vision (ICCV)*, pages 2992–2992, 2015. 3

[27] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *European Conference on Computer Vision (ECCV)*, 2012. 2, 3, 5, 7

[28] Thomas Nestmeyer and Peter V Gehler. Reflectance adaptive filtering improves intrinsic image estimation. In *IEEE*

Conference on Computer Vision and Pattern Recognition (CVPR), page 4, 2017. 2, 3, 6

[29] Geoffrey Oxholm and Ko Nishino. Shape and reflectance from natural illumination. In *European Conference on Computer Vision (ECCV)*, pages 528–541. Springer, 2012. 2

[30] Emmanuel Prados and Olivier Faugeras. Shape from shading. In *Handbook of mathematical models in computer vision*, pages 375–388. Springer, 2006. 2

[31] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention (MICCAI)*, pages 234–241. Springer, 2015. 4

[32] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D. Castillo, and David W. Jacobs. Sfsnet: Learning shape, refectance and illuminance of faces in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 3, 4, 5, 7

[33] Evan Shelhamer, Jonathan T Barron, and Trevor Darrell. Scene intrinsics and depth from a single image. In *International Conference on Computer Vision, Workshops (ICCV-W)*, pages 37–44, 2015. 2

[34] Jian Shi, Yue Dong, Hao Su, and X Yu Stella. Learning non-lambertian object intrinsics across shapenet categories. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5844–5853, 2017. 2, 3

[35] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2930–2937, 2013. 7

[36] Zhixin Shu, Ersin Yumer, Sunil Hadap, Kalyan Sunkavalli, Eli Shechtman, and Dimitris Samaras. Neural face editing with intrinsic image disentangling. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5444–5453, 2017. 2, 4

[37] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 3, 4, 5, 6

[38] Tatsunori Taniai and Takanori Maehara. Neural inverse rendering for general reflectance photometric stereo. In *Intl. Conf. on Machine Learning (ICML)*, pages 20–28, 2017. 2

[39] Marshall F Tappen, William T Freeman, and Edward H Adelson. Recovering intrinsic images from a single image. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1367–1374, 2003. 2

[40] B Tunwattanapong and P Debevec. Interactive image-based relighting with spatially-varying lights. In *ACM Transactions on Graphics (SIGGRAPH)*, 2009. 1

[41] Tuanfeng Wang, Tobias Ritschel, and Niloy Mitra. Joint material and illumination estimation from photo sets in the wild. In *International Conference on 3D Vision (3DV)*, pages 22–31, 2018. 2

[42] Henrique Weber, Donald Prévost, and Jean-François Lalonde. Learning to estimate indoor lighting from 3d objects. In *International Conference on 3D Vision (3DV)*. 7

[43] Zexiang Xu, Kalyan Sunkavalli, Sunil Hadap, and Ravi Ramamoorthi. Deep image-based relighting from optimal sparse samples. *ACM Transactions on Graphics (TOG)*, 37(4):126, 2018. 1

[44] Edward Zhang, Michael F Cohen, and Brian Curless. Emptying, refurnishing, and relighting indoor spaces. *ACM Transactions on Graphics (TOG)*, 35(6):174, 2016. 2

[45] Edward Zhang, Michael F Cohen, and Brian Curless. Discovering point lights with intensity distance fields. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6635–6643, 2018. 2

[46] Yinda Zhang, Shuran Song, Ersin Yumer, Manolis Savva, Joon-Young Lee, Hailin Jin, and Thomas Funkhouser. Physically-based rendering for indoor scene understanding using convolutional neural networks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2, 3, 4, 5, 6, 7

[47] Hao Zhou, Jin Sun, Yaser Yacoob, and David W Jacobs. Label denoising adversarial network (LDAN) for inverse lighting of face images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[48] Tinghui Zhou, Philipp Krahenbuhl, and Alexei A Efros. Learning data-driven reflectance priors for intrinsic image decomposition. In *International Conference on Computer Vision (ICCV)*, pages 3469–3477, 2015. 2, 3, 5, 6

[49] Wei Zhuo, Mathieu Salzmann, Xuming He, and Miaomiao Liu. Indoor scene structure analysis for single image depth estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 614–622, 2015. 2