GLoSH: Global-Local Spherical Harmonics for Intrinsic Image Decomposition

Hao Zhou[1] * Xiang Yu[2] David W. Jacobs[1]
[1] University of Maryland, College Park, MD, USA
[2] NEC Laboratories America
[1]{hzhou, djacobs}@cs.umd.edu [2]xiangyu@nec-labs.com

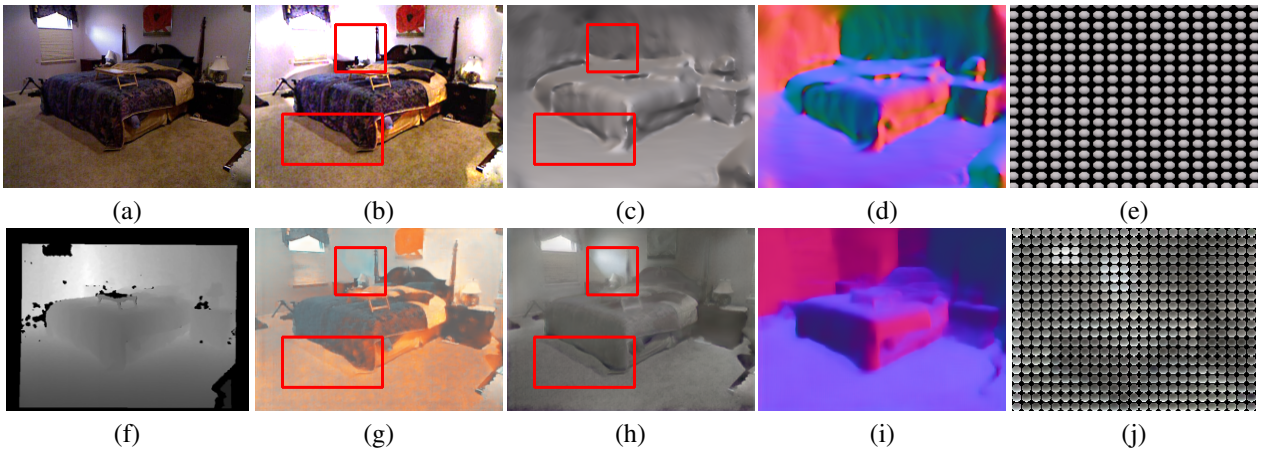


Figure 1: Top row: result of [2]. [2] takes an **RGB-D** image as input and predicts: (b) reflectance, (c) shading, (d) normal, (e) lighting. Bottom row: result of our method. It takes an **RGB** image as input and predicts: (g) reflectance, (h) shading, (i) normal and (j) lighting. The red boxes show that our algorithm correctly attributes cast shadows and highlights to shading while [2] incorrectly attributes them to reflectance. Our lighting (j) captures local lighting variation better than (e) from [2].

Abstract

Traditional intrinsic image decomposition focuses on decomposing images into reflectance and shading, leaving surfaces normals and lighting entangled in shading. In this work, we propose a Global-Local Spherical Harmonics (GLoSH) lighting model to improve the lighting component, and jointly predict reflectance and surface normals. The global SH models the holistic lighting while local SH account for the spatial variation of lighting. Also, a novel non-negative lighting constraint is proposed to encourage the estimated SH to be physically meaningful. To seamlessly reflect the GLoSH model, we design a coarse-to-fine network structure. The coarse network predicts global SH, reflectance and normals, and the fine network predicts their local residuals. Lacking labels for reflectance and lighting, we apply synthetic data for model pre-training and fine-tune the model with real data in a self-supervised way. Compared to the state-of-the-art methods only targeting normals or reflectance and shading, our method recovers all components and achieves consistently better results on three real datasets, IIW, SAW and NYUv2.

*Hao Zhou is currently at Amazon AWS.

1. Introduction

Understanding the physical world that produces an image is a core problem in computer vision. [4] first proposed to estimate the intrinsic scene characteristics from images, including range, orientation, reflectance and incident lighting. This is a notoriously difficult inverse problem as it is highly under-constrained. Moreover, we lack models of the physical components of the problem, such as lighting, that are both accurate and easy to use. Early works start with investigating the reflectance, shape and illumination of a single object [1, 3], as the lighting for a single object is easier to model, for instance, by using a single set of low dimensional Spherical Harmonics [5, 31]. The lighting of a natural scene, however, is much more complicated due to its spatial variation caused by shadow, inter-reflection and the presence of light sources in the scene. As a result, most works that address scenes have lumped normal and lighting together as shading, and try to recover that, known as intrinsic image decomposition.

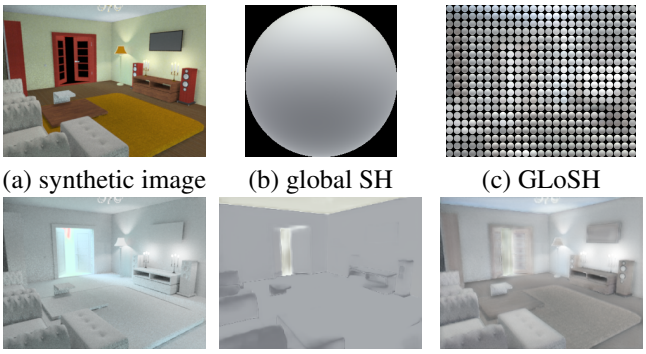
In this paper, we propose a new representation of lighting for scenes, which allows us to disentangle lighting and surface normals, while also recovering reflectance. One way to model lighting is Spherical Harmonics (SH) [5, 31], which approximates the lighting with 9 low frequency components. While this works well for modeling the lighting of small objects, such as faces [5, 42, 33], such a global lighting cannot capture the spatially varying lighting in a complex scene, as shown in Figure 2 (e). Allowing independent lighting in each pixel, however, creates too many degrees of freedom and would allow lighting variation alone to explain the image.

To overcome the problem, we propose a Global-Local Spherical Harmonics (GLoSH) lighting model. Our global SH represents the holistic lighting of the entire scene. On top of it, the local SH, produced by the sum of global SH and local residual SH, account for the spatial variation of the lighting. An L_2 regularization on the local residual SH limits the effects of over-parameterization. Figure 2 (c) shows our GLoSH and Figure 2 (f) shows the reconstructed shading, which is much closer to ground truth than only using global SH.

Spherical Harmonics with arbitrary coefficients would represent lighting in a physically unrealistic way, if the lighting is negative in some directions. Nevertheless, enforcing non-negative SH lighting is not trivial. Existing methods either introduce many more parameters to constrain non-negative lighting [5] or require solving a semi-definite programming problem [36], which is difficult to directly incorporate with deep networks. In this work, we propose to sample the intensity of the lighting uniformly distributed on a sphere generated from the predicted SH. A non-negative loss is then defined on the sampled lighting. Our non-negative constraint is only applied to global SH, because practically the local residual SH regularized by L_2 are not likely to change the sign of the lighting.

We apply a CNN to achieve an end-to-end coarse-to-fine solution. Training deep CNNs requires huge amounts of data and ground truth labels, and labeling images for reflectance and lighting is extremely difficult. Intrinsic Images in the Wild (IIW) [7] labels the relative darkness of the reflectance from pairs of pixels. Shading Annotations in the Wild (SAW) [22] labels constant shading regions, shadow boundaries and depth/normal discontinuities. However, these datasets only provide sparse labels and a limited number of images. Inspired by recent success of synthetic data on computer vision applications, we propose to use the synthesized SUNCG dataset [37], in which ground truth reflectance, normal and shading can be easily determined, to pre-train the models. The pre-trained model is then further trained with real data in a self-supervised way.

To sum up, we propose a GLoSH lighting model, and apply a coarse-to-fine CNN structure to predict GLoSH together with reflectance and normal. The synthetic data pre-training and self-supervised training with real data lead to state-of-the-art performance across three real scene datasets, IIW, SAW and NYUv2. The contributions of our work are as the following.



(a) synthetic image (b) global SH (c) GLoSH

(d) GT shading (e) shading w.r.t. (b) (f) shading w.r.t. (c)

Figure 2: Visualization of global SH modeling (b) and its reconstructed shading (e), comparing to our GLoSH (c) and its reconstructed shading (f). With GLoSH, clearly our method generates the shading much closer to ground truth.

- We propose a GLoSH lighting model with global and local SH, and a novel non-negative constraint to estimate physically realistic lighting.
- To the best of our knowledge, under a single RGB image setting, we are the first to apply CNNs to jointly estimate reflectance, normal and lighting.
- We propose a coarse-to-fine network that is compatible with our proposed global-local lighting model.
- Our method achieves the best results on IIW reflectance, the second best on SAW shading, and strongly competitive performance on NYUv2 normal. Notice that the state-of-the-art methods only focus on one or two components, while our method jointly estimates reflectance, normal and lighting.

2. Related Work

Intrinsic Image Characteristics. We categorize the literature into two main streams: single object based and natural scene based methods. Researchers have long been studying the estimation of intrinsic image characteristics for a single object. For example, shape from shading [39, 11] focuses on recovering the shape assuming illumination and reflectance are known. Photometric stereo [38] estimates geometry from multiple images assuming known lighting. Recent progress in photometric stereo [1] can estimate geometry and lighting up to a bas-relief transformation [6]. [15, 35, 25] proposed to decompose a single object image into its reflectance and shading. [3] and [18] proposed to jointly estimate reflectance, shape and lighting from a sin-

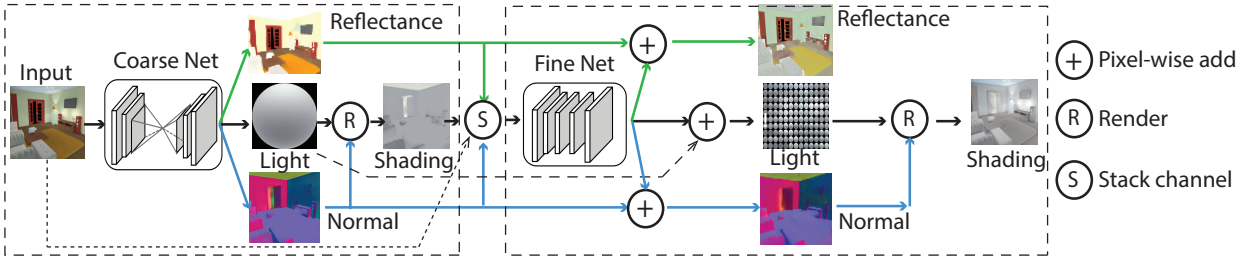


Figure 3: Our coarse-to-fine network structure. The coarse net predicts the first level reflectance, lighting and surface normal, of which the latter two further form the shading. The fine net takes the previous stacked output as input and predicts the residual of reflectance, lighting and surface normal. The final reflectance, lighting and normal are recovered by adding the predicted residual with the first level results.

gle object image.

Estimating a natural scene is more difficult due to more complicated geometry and lighting. Recent studies [32, 12] show the capability of accurately estimating the scene geometry thanks to large scale training data and the success of deep learning. Some advanced methods [9, 19, 2, 34, 7, 8] proposed optimization based approaches to decompose an image into reflectance and shading, where [9, 19, 2] require depth to be known. Most recent work [43, 29, 26, 27, 10, 20, 44, 24, 23, 13] applies deep Convolutional Neural Networks (CNN) on this task and achieve impressive performance. [23] proposed to render realistic synthetic data and then use it to train the deep models and adapt to the real dataset. Our work follows a similar idea. However, we not only estimate reflectance and shading, but also further decompose the shading into normal and lighting.

Barron and Malik [2] first proposed to estimate reflectance, depth/normal and lighting with RGB-D images. They model the lighting at each pixel as a linear combination of eight sets of Spherical Harmonics. In contrast, we jointly estimate reflectance, surface normals and lighting from a single RGB image without depth, which is a much harder problem. Moreover, we propose our Global-Local SH (GloSH) with a coarse-to-fine neural network to represent the lighting for each pixel, which accounts for not only the holistic lighting but also the local lighting variations.

Non-negative Spherical Harmonics. While using Spherical Harmonics lighting, one challenge is how to enforce lighting to be non-negative. [5] proposed to represent lighting using a non-negative linear combination of delta functions to solve this problem. One drawback of this method is that to have an accurate representation, a lot of delta functions are needed. [36] proved that the Toeplitz matrix of a non-negative SH is positive semi-definite. They proposed to solve a semi-definite programming (SDP) problem to enforce non-negative lighting. However, the SDP constraint is not obviously tractable to incorporate with deep training. In contrast, we formulate a non-negative lighting loss by sampling hundreds of points on a predicted lighting sphere, which is computationally efficient and fits into the network training smoothly.

3. Reflectance, Normal and Shading from a Single RGB Image

Intrinsic image decomposition assumes an image $\mathbf{I}$ to be the product of reflectance $\mathbf{R}$ and shading $\mathbf{S}$, *i.e.* $\mathbf{I} = \mathbf{R} \odot \mathbf{S}$, where $\odot$ represents an element-wise product. Most research studies focus on decomposing an image $\mathbf{I}$ into $\mathbf{R}$ and $\mathbf{S}$, where geometry and lighting remain entangled in shading. In our work, we propose to further decompose shading $\mathbf{S}$ into surface normal (*i.e.* geometry) $\mathbf{N}$ and lighting $\mathbf{L}$. Assuming $\mathbf{S} = \Psi(\mathbf{N}, \mathbf{L})$, an image $\mathbf{I}$ can be represented as

$$\mathbf{I} = \mathbf{R} \odot \Psi(\mathbf{N}, \mathbf{L}), \tag{1}$$

Ψ is a rendering function. Our target is to estimate $\mathbf{R}$, $\mathbf{N}$ and $\mathbf{L}$ given a single image $\mathbf{I}$.

3.1. GLoSH Lighting Modeling

While a single, global set of low-dimensional SH have been used to represent lighting of objects, this would be unable to capture the complex lighting conditions of a scene. On the other hand, estimating SH for each pixel easily falls into over-parameterization. We propose a neural network based Global-Local Spherical Harmonics (GLoSH) model, where global SH serves as the low frequency approximation of the lighting, and local residual SH accounts for spatial variation. A coarse-to-fine neural structure is designed to exactly execute the global and local lighting modeling.

3.1.1 Global and local Spherical Harmonics

Following [5, 31], we propose to use SH up to the second order, resulting in a 9 dimensional SH for each color channel. Denote the global SH as $\mathbf{L}_c \in \mathbf{R}^9$. $\mathbf{L}_c$ is predicted from our coarse level network

As revealed in Figure 2, based only on the global SH $\mathbf{L}_c$, the shading is far from satisfactory, lacking much spatial variation. To better model the spatial variation of the lighting, we predict local residual SH for each pixel in a fine

level network. Our local SH is then formulated as global SH with the local residual SH:

$$\mathbf{L}_f = \mathbf{L}_c + \delta\mathbf{L}_f, \tag{2}$$

where $\delta\mathbf{L}_f$ represent the local residual SH predicted by a fine scale network.

3.1.2 Non-negative Constraints on SH

Physically realistic lighting requires non-negative SH lighting, which previous work [3, 2] does not properly consider. To enforce the non-negative SH lighting, we propose a simple yet effective constraint on SH. According to [5], given a SH coefficient $\mathbf{L}_c$, the lighting intensity at a direction (θ, ϕ) is a function of $\mathbf{L}_c$, *i.e.*, $f_{\mathbf{L}}(\mathbf{L}_c, \theta, \phi)$. A non-negative lighting means $f_{\mathbf{L}}(\mathbf{L}_c, \theta, \phi) \geq 0, \forall 0 \leq \theta \leq \pi, 0 \leq \phi \leq 2\pi$. Based on this, we uniformly sample the value of the function $f_{\mathbf{L}}$ on a unit sphere and constrain all the sampled values to be non-negative. The non-negative loss function is thus defined as

$$\mathcal{L}_{Lc} = \frac{1}{K}\sum_{i=1}^{K} \min(0, f_{\mathbf{L}}(\mathbf{L}_c, \theta_i, \phi_i))^2, \tag{3}$$

$K = 6414$ is the number of directions sampled from the sphere. We apply this non-negative constraint to global SH. We further apply the L_2 regularization over the local residual SH:

$$\mathcal{L}_{Lf} = \|\delta\mathbf{L}_f\|_2^2. \tag{4}$$

This regularization penalizes their L2 norm, encouraging the local lighting to not vary too much from the global lighting.

Our experiments demonstrate that Equation 3 and Equation 4 almost always result in the non-negative lighting for local SH.

3.2. Coarse-to-fine Network Structure

To exactly match the proposed GLoSH lighting modeling, we design a coarse-to-fine network structure shown in Figure 3. The coarse network is defined as an hourglass network [30]. It takes an image $\mathbf{x} \in \mathbb{R}^{64\times64\times3}$ as input and predicts reflectance $\mathbf{R}_c \in \mathbb{R}^{64\times64\times3}$, normal $\mathbf{N}_c \in \mathbb{R}^{64\times64}$, and global SH $\mathbf{L}_c \in \mathbb{R}^9$. Shading $\mathbf{S}_c \in \mathbb{R}^{64\times64\times3}$ can be constructed by a simple rendering function.

$$\mathbf{S}_c = \Psi(\mathbf{N}_c, \mathbf{L}_c) \tag{5}$$

The fine scale network is designed with fully convolutional structures. It takes $\mathbf{x} \in \mathbb{R}^{128\times128\times3}$, upsampled $\mathbf{R}_c \in \mathbb{R}^{128\times128\times3}, \mathbf{N}_c \in \mathbb{R}^{128\times128}, \mathbf{S}_c \in \mathbb{R}^{128\times128\times3}$ as input and predicts residual maps. The recovered local reflectance, normal and local SH are:

$$\begin{aligned}
\mathbf{R}_f &= \mathbf{R}_c + \Phi_f^{\mathbf{R}_f}(\mathbf{R}_c, \mathbf{N}_c, \mathbf{L}_c), \\
\mathbf{N}_f &= \mathbf{N}_c + \Phi_f^{\mathbf{N}_f}(\mathbf{R}_c, \mathbf{N}_c, \mathbf{L}_c), \\
\mathbf{L}_f &= \mathbf{L}_c + \Phi_f^{\mathbf{L}_f}(\mathbf{R}_c, \mathbf{N}_c, \mathbf{L}_c),
\end{aligned} \tag{6}$$

where, $\Phi_f^{\mathbf{R}_f}$, $\Phi_f^{\mathbf{N}_f}$ and $\Phi_f^{\mathbf{L}_f}$ represent the fine level network for reflectance, normal and lighting respectively. The fine scale shading is calculated by $\mathbf{S}_f = \Psi(\mathbf{N}_f, \mathbf{L}_f)$. The fine scale network structure can be recurrently applied to a finer scale. Our full model is defined to have three scales, which can predict reflectance, normal and lighting with resolution 256×256. Please refer to the supplementary materials for more detail.

3.3. Supervision on Training

It is difficult to obtain dense accurate ground truth annotation for reflectance, normal and lighting. We thus leverage the rendered synthetic data for the supervised pre-training. The pre-trained network is then fine-tuned using sparsely annotated real data (IIW [7], SAW [22] and NYUv2 [28]) in a self-supervised way, *i.e.*, applying the pre-trained model to provide pseudo ground truth labels for fine-tuning.

3.3.1 Reflectance

In the pre-training stage, we directly apply the ground truth reflectance to guide the training in a fully supervised way, where L_1 loss is applied as shown in Equation 7.

$$\mathcal{L}_{R1} = \|\mathbf{R} - \mathbf{R}^*\|_1 + \|\nabla\mathbf{R} - \nabla\mathbf{R}^*\|_1. \tag{7}$$

$\mathbf{R}$ is the predicted reflectance and $\mathbf{R}^*$ is the corresponding ground truth. Moreover, similar to [23], we add supervision to the gradient of the reflectance to encourage the predicted reflectance to be piece-wise smooth.

For real data, there is no dense annotation for either reflectance, normal or lighting. Instead, IIW [7] provides sparse ordinal reflectance judgments. Given a pair of reflectances $\mathbf{R}_1$ and $\mathbf{R}_2$, the label indicates whether $\mathbf{R}_1$ is darker than (lighter than or equal to) $\mathbf{R}_2$ (demoted as $J = 1$, $J = -1$ and $J = 0$ respectively) with a confidence score w. We use the WHDR hinge loss proposed in [29] as the loss for reflectance in real images:

$$\mathcal{L}_R(\mathbf{R}_1, \mathbf{R}_2, J) = \tag{8}$$
$$w \begin{cases} \max\left(0, \frac{R_1}{R_2} - \frac{1}{1+\delta+\xi}\right) & \text{if} \quad J = 1 \\ \max\left(0, \begin{cases} \frac{1}{1+\delta-\xi} - \frac{R_1}{R_2} \\ \frac{R_1}{R_2} - (1+\delta-\xi) \end{cases}\right) & \text{if} \quad J = 0 \\ \max\left(0, (1+\delta+\xi) - \frac{R_1}{R_2}\right) & \text{if} \quad J = -1 \end{cases}$$

We set $\delta = 0.12$ and $\xi = 0.08$ during training as in [29]. Notice that the above loss is not symmetric, *i.e.*, $\mathcal{L}_{R_i}(\mathbf{R}_1, \mathbf{R}_2, J) \neq \mathcal{L}_R(\mathbf{R}_2, \mathbf{R}_1, -J)$. We thus adapt the above loss and define the modified WHDR loss as:

$$\mathcal{L}_{R2} = \mathcal{L}_R(\mathbf{R}_1, \mathbf{R}_2, J) + \mathcal{L}_R(\mathbf{R}_2, \mathbf{R}_1, -J) \tag{9}$$

3.3.2 Normal

The ground truth normal for synthetic data and part of the real data (NYUv2) are available. For those data in which ground truth normals are available, we define the loss as

$$\mathcal{L}_N = -\mathbf{N}^T\mathbf{N}^* + \|\nabla\mathbf{N} - \nabla\mathbf{N}^*\|_1 \tag{10}$$

Similar to reflectance regularization in Equation 7, we further apply the first order derivative smoothness term to encourage the normal to be piece-wise continuous.

3.3.3 Shading

There is no supervision for lighting. The non-negative constraint and the L_2 regularization are all unsupervised losses. Applying rendering to generate shading $\mathbf{S} = \Psi(\mathbf{N}, \mathbf{L})$ from normal and lighting, we use the supervision on shading and normal discussed in Sec. 3.3.2 to indirectly supervise the lighting. The supervised signal for shading is similar to that of reflectance:

$$\mathcal{L}_{S1} = \|\mathbf{S} - \mathbf{S}^*\|_1 + \|\nabla\mathbf{S} - \nabla\mathbf{S}^*\|_1 \tag{11}$$

where $\mathbf{S}$ and $\mathbf{S}^*$ are predicted shading and its ground truth.

For real images, SAW [22] provides annotation for smooth shading regions and shadow boundaries. We thus apply the same loss as in [23] for the shading:

$$\mathcal{L}_{S2} = \lambda_{cs}\mathcal{L}_{constant-shading} + \mathcal{L}_{shadow} \tag{12}$$

where $\lambda_{cs} = 10$ and $\mathcal{L}_{constant-shading}$ and $\mathcal{L}_{shadow}$ are the loss for constant shading region and shadow boundary defined in [23].

4. Implementation Details

Pre-training on Synthetic Data: We first train our network using the SUNCG dataset with synthesized ground truth normal, reflectance, and shading. The loss to train our network on synthetic data is

$$\mathcal{L}_s = \lambda_{sR}\mathcal{L}_{R1} + \lambda_{sS}\mathcal{L}_{S1} + \lambda_{sN}\mathcal{L}_N + \lambda_{Lc}\mathcal{L}_{Lc} + \lambda_{Lf}\mathcal{L}_{Lf} \tag{13}$$

where $\mathcal{L}_{R1}$, $\mathcal{L}_{S1}$, $\mathcal{L}_N$, $\mathcal{L}_{Lc}$ and $\mathcal{L}_{Lf}$ are losses for reflectance, shading, normal, global and local residual lighting defined above, and λ_{sR}, λ_{sS}, λ_{sN}, λ_{Lc} and λ_{Lf} are their corresponding weights. We set $\lambda_{sR} = \lambda_{sS} = \lambda_{sN} = \lambda_{Lc} = 1$ and $\lambda_{Lf} = 0.2$. Our coarse-to-fine network is trained step by step using the Adam [21] optimizer with initial learning rate 0.001 and weight decay 0.

Fine-tuning on Real Data: Due to the lack of annotation from real datasets, we use the rendered SUNCG dataset as supervision, with the loss denoted as $\mathcal{L}_r^{cg}$. In addition, we apply our network trained on synthetic data to predict reflectance, shading and normal of real images and use the results as pseudo supervision (self-supervision), with the loss denoted as $\mathcal{L}_r^{ss}$.

$$\begin{aligned}\mathcal{L}_r^{cg} &= \lambda_{sR}^{cg}\mathcal{L}_{R1} + \lambda_{sS}^{cg}\mathcal{L}_{S1} + \lambda_{sN}^{cg}\mathcal{L}_N + \lambda_{Lc}^{cg}\mathcal{L}_{Lc} + \lambda_{Lf}^{cg}\mathcal{L}_{Lf},\\ \mathcal{L}_r^{ss} &= \lambda_{rR}^{ss}\mathcal{L}_{R1} + \lambda_{rS}^{ss}\mathcal{L}_{S1} + \lambda_{rN}^{ss}\mathcal{L}_N + \lambda_{Lc}^{ss}\mathcal{L}_{Lc} + \lambda_{Lf}^{ss}\mathcal{L}_{Lf}\end{aligned} \tag{14}$$

where we set $\lambda_{sR}^{cg} = \lambda_{sS}^{cg} = \lambda_{Lc}^{cg} = \lambda_{Lf}^{cg} = 1$, $\lambda_{sN}^{cg} = 10$, $\lambda_{rS}^{ss} = \lambda_{rN}^{ss} = 5$, $\lambda_{Lf}^{ss} = \lambda_{Lc}^{ss} = 1$ and $\lambda_{rR}^{ss} = 0.1$. Our loss defined on the annotation and ground truth of IIW, SAW and NYUv2 is:

$$\mathcal{L}_r^o = \lambda_{rR}^o\mathcal{L}_{R2} + \lambda_{rS}^o\mathcal{L}_{S2} + \lambda_{rN}^o\mathcal{L}_N \tag{15}$$

where $\lambda_{Lc}^o = \lambda_{rN}^o = 10$, $\lambda_{rS}^o = 1$. Inspired by [7], we introduce the L_2 regularization to achieve a reasonable color for reflectance.

$$\mathcal{L}_r^c = \|\frac{\mathbf{R}}{\frac{1}{3}\sum_c \mathbf{R}^c} - \frac{\mathbf{I}}{\frac{1}{3}\sum_c \mathbf{I}^c}\|_1 \tag{16}$$

where $\mathbf{R}$ and $\mathbf{I}$ are predicted reflectance and input image, and $\mathbf{R}^c$ and $\mathbf{I}^c$, $c \in \{R, G, B\}$ denote the color channel of $\mathbf{R}$ and $\mathbf{I}$. Importantly, a reconstruction loss is further introduced to guarantee that the predicted reflectance, normal and lighting preserve the input's characteristics.

$$\mathcal{L}_r^{rc} = \|\mathbf{I}_i - \mathbf{R}_i \odot \mathbf{S}_i\|_2 \tag{17}$$

The overall loss that we apply to fine-tune our network on real images is:

$$\mathcal{L}_r = \mathcal{L}_r^{cg} + \mathcal{L}_r^{ss} + \mathcal{L}_r^o + \mathcal{L}_r^c + \lambda_r^{rc}\mathcal{L}_r^{rc} \tag{18}$$

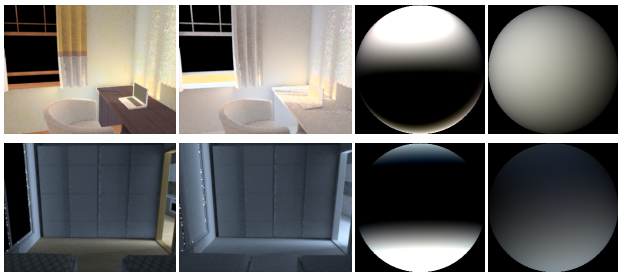
where $\lambda_r^{rc} = 0.1$. The coarse-to-fine network is fine-tuned scale by scale. The Adam optimizer with learning rate 0.0005 and weight decay 0.00001 is used for fine-tuning.

5. Experiments

In this section, we introduce the synthetic dataset that we create for pre-training and the public real datasets. Then we compare to Barron and Malik [2], who first proposed to predict reflectance, normal and lighting from an RGB-D image. Further, we compare to the state-of-the-art intrinsic image decomposition methods to indicate the overall advantage of our method. An ablation study is then carried out to demonstrate the contribution of each of our proposed modules.

5.1. Datasets

Synthetic Dataset: we make use of the SUNCG dataset [40] to generate synthetic data. It contains $568,793$ images rendered using Mitsuba [17] and their corresponding ground truth surface normals, depths, semantic labels and object boundaries. Since our task also requires ground truth reflectance and shading, we re-render $58,949$ images of SUNCG using the multi-channel renderer of Mitsuba. We further split the images into a training set of $51,507$ images and a validation set of $7,442$ images. Instead of directly rendering images, we render shading by setting all the materials to diffuse and the reflectance to be 1. Then the image



(a) (b) (c) (d)

Figure 4: (a) synthetic images, (b) shading images, (c) and (d) are lighting predicted by training the network without and with non-negative constraint respectively.

Table 1: SH lighting Evaluation on SUNCG synthetic data.

	[2]	GLoSH SUNCG	GLoSH SUNCG + real
MSE	0.098	0.038	**0.032**

is $\mathbf{I} = \mathbf{R} \odot \mathbf{S}$. Rendering in this way has two main advantages: (1) The generated images strictly follow the assumption of intrinsic image decomposition. (2) The pixel value of ground truth shading has bounded range which makes data preparation easier. Though the rendered images do not contain non-diffuse effects of the material, our experiments show that this does not degrade the performance.

Public Real Datasets: we use IIW [7], SAW [22] and NYUv2 [28] as real data for training and testing. More specifically, SAW is a combination of IIW and NYUv2 (3761 images from IIW and 381 images from NYUv2 with ground truth normals). The real dataset we use is the same as [23] in addition to ground truth normals from NYUv2. We strictly follow the train/val/test splitting strategy of [23].

5.2. Spherical Harmonics Lighting Evaluation

Quantitative comparison to [2]. We compare to [2] as they also propose a lighting model to jointly predict reflectance, normal and lighting of a natural scene. Notice that [2] uses RGB-D images, which simplifies the problem.

Lighting for real data is hard to obtain. We instead evaluate the shading from the SUNCG synthetic data by fixing the surface normal from ground truth, at which we can indirectly evaluate the SH lighting. We calculate the per-pixel Mean squared error (MSE) of the reconstructed shading w.r.t. ground truth shading and show the results in Table 1. Our method shows a significant advantage over [2] and the real data self-supervision provides a further performance boost. We also evaluate the shading of [2] on NYUv2 dataset using the AP challenge metric proposed by [23]. They achieve $\mathbf{90.38}\%$ shading accuracy, while under the same setup, our method achieves $\mathbf{95.43}\%$. We believe all these results show that the proposed method can predict much more accurate lighting than [2].

Table 2: Surface normal evaluation on NYUv2. Average (Avg.) and Median (Med.) show the average and median angular error, smaller values are the better. $11.25°$, $22.5°$ and $30°$ shows the percentage of normals with angular error smaller than $11.25°$, $22.5°$ and $30°$, higher values are the better.

Method	Avg. (°)↓	Med. (°)↓	11.25° ↑	22.5° ↑	30° ↑
[40]	**27.90**	21.29	26.76	52.21	**63.75**
Ours	28.63	**21.05**	**27.68**	**52.42**	62.87

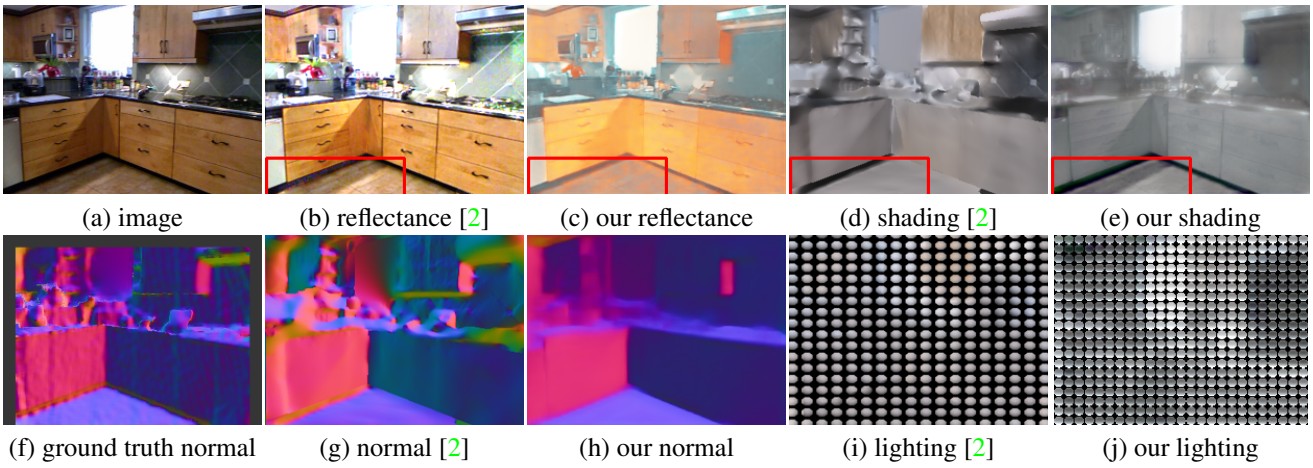
Qualitative comparison to [2]. Figure 5 compares their visual results with ours. The red rectangles in reflectance and shading images show that [2] mistakenly decomposes cast shadow into reflectance instead of shading. We believe the limited number of SH basis in their method prevents them from modeling the spatial variation of the lighting well, resulting in a lack of ability to model cast shadow.

Non-negative lighting: [36] proved that a SH represents non-negative lighting if its Toeplitz matrix is positive semi-definite. We use their proposed method to evaluate the effectiveness of our non-negative constraint. We train our coarse scale network with and without the proposed non-negative constraint, *i.e.*, Equation (3), and then test on the validation set of our synthetic SUNCG data. Without the proposed non-negative constraint, the percentage of global SH that represents negative lighting is $\mathbf{13.39}\%$. It drastically decreases to $\mathbf{1.09}\%$ with this constraint. Figure 4 visualizes the predicted lighting with and without the non-negative constraint. After fine-tuning on real data, the global SH that represents negative lighting is reduced to $\mathbf{0}\%$ and there is only one image that contains negative local lighting.

5.3. Intrinsic Image Decomposition

Model trained on synthetic data. We evaluate our network trained using synthetic data on IIW, SAW and NYUv2. For reflectance on IIW, we use the WHDR metric proposed in [7], which computes the weighted error of the predicted reflectance with human annotation. The challenge average precision (AP) proposed by [23] is used to evaluate the predicted shading. It computes the average precision of classification for constant shading regions and shadow boundaries. Table 3 (a) compares our trained network with [23] on IIW and SAW dataset. It shows that our proposed method are closely comparable to [23] on IIW and much better than [23] on SAW when trained on SUNCG dataset.

[23] claimed that the dataset they provided (denoted as CGI) has a smaller domain gap with real data compared with SUNCG. For a sanity check, we train our coarse network using CGI and achieve WHDR 37.98, while the WHDR of our coarse network trained on SUNCG is 28.20.



(a) image (b) reflectance [2] (c) our reflectance (d) shading [2] (e) our shading

(f) ground truth normal (g) normal [2] (h) our normal (i) lighting [2] (j) our lighting

Figure 5: Comparison with [2]. Red rectangle shows that our method can correct decompose cast shadows into shading while [2] cannot. Due to space limits, please refer to supplementary materials for more results.

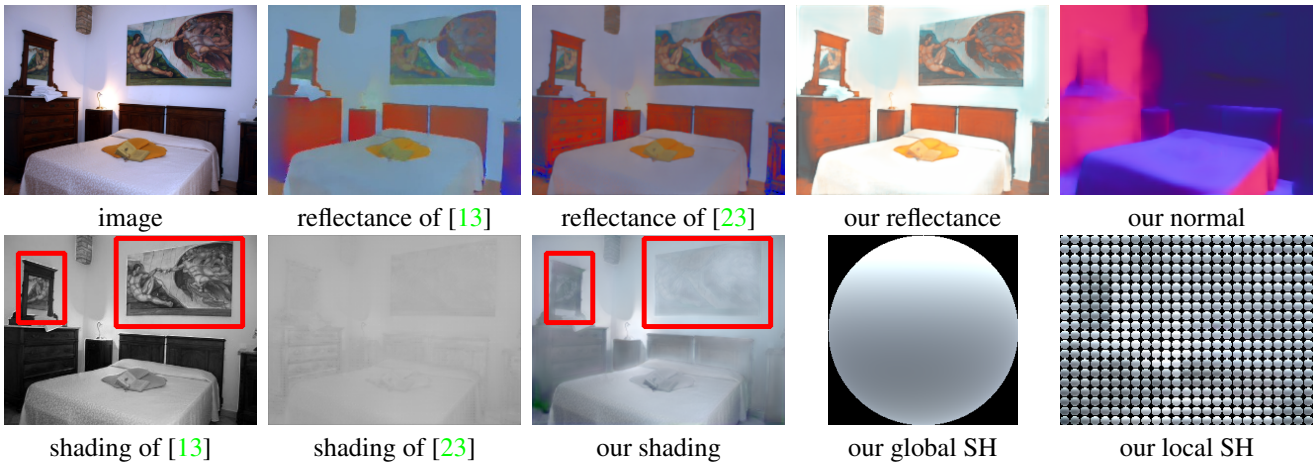


image reflectance of [13] reflectance of [23] our reflectance our normal

shading of [13] shading of [23] our shading our global SH our local SH

Figure 6: Comparison with state-of-the-art intrinsic image decomposition methods. Note that although [23] achieves the best AP score on shading, the generated shading image is of very low contrast. The red rectangle shows the shading of [13] suffers seriously from the reflectance bleeding problem. Due to space limits, please refer to supplementary materials for more results.

We do not see the advantage of using CGI data for training and thus we train our network using SUNCG dataset.

Model fine-tuned on real data. Table 3 (b) compares our method with some start-of-the-art methods on IIW and SAW. Our method achieves the best performance on IIW and second best on SAW.

[13] demonstrated that by incorporating the guided filter into the training of their network, they can achieve a WHDR of 14.5% which is the state-of-the-art result. By applying a guided filter to our model as suggested by [29], we can achieve 14.6%; which is closely comparable to this result. However, the challenge AP for the shading of [13] on IIW dataset [1] is $\mathbf{85.77}\%$. Under the same setting, we achieve $\mathbf{97.08}\%$, a more than 10% improvement.

Besides reflectance and shading, Table 2 shows that the normal predicted by our model achieves strongly competitive results with [40], when trained on SUNCG synthetic data and evaluated on NYUv2. We further fine-tuned the models with limited real data (381 images with surface normal ground truth), and achieved $25.57°$ average angular error, close to [40] $21.74°$.

Visual comparison. We visualize the shading predicted by [13] in Figure 6 (c). It shows that the shading images of [13] still retain the effect of reflectance. Although [23] achieves the best performance on SAW, Figure 6 (e) shows that their predicted shading images are of low contrast. That is, the quality of the shading image is low. Across the compared

[1] Images are provided by the authors.

Table 3: Reflectance evaluation on IIW and shading evaluation on SAW. For WHDR, lower value (↓) is better, for AP, higher value is better(↑).

			IIW	SAW
	Method	Dataset	WHDR (%)↓	AP (%)↑
a	Li [23]	SUNCG	**26.1**	87.09
	Proposed	SUNCG	26.8	**92.40**
b	Grosse [16]	-	26.9	85.26
	Garces [14]	-	24.8	92.39
	Zhao [41]	-	23.8	89.72
	Bi [8]	-	17.7	-
	Bell [7]	-	20.6	92.18
	Zhou [43]	IIW	19.9	86.34
	[29]	IIW	19.5	89.94
	Fan [13]	IIW	15.4	-
	Li [23]	CGI + *real*	15.5	**96.57**
	proposed	SUNCG + *real*	**15.2**	95.01

Table 4: Ablation study on loss, without synthetic SUNCG data, and the coarse-to-fine scales, evaluated on IIW reflectance, SAW shading and NYUv2 surface normal.

	IIW	SAW	NYUv2
Method	WHDR (%) ↓	AP (%) ↑	Mean Error (°) ↓
w/o SUNCG	17.82	88.52	35.14
w/o $\mathcal{L}_r^{ss}$	15.50	**95.79**	25.93
w/o $\mathcal{L}_{R2}$	15.34	91.89	25.96
scale1	18.70	90.35	26.68
scale1+scale2	16.62	94.98	25.59
full	**15.20**	95.01	**25.57**

methods, our method achieves relatively better visual quality on both reflectance and shading.

To conclude, our GLoSH achieves consistently better results compared to state-of-the-art methods trained on both synthetic data and fine-tuned on real data, across the tasks of estimating reflectance, normal, shading and lighting. We believe this also indicates the effectiveness of the proposed coarse-to-fine network structure.

5.4. Ablation Study

Without synthetic data. Synthetic data is very important for the proposed method. Table 4 "w/o SUNCG" shows the WHDR on IIW, average precision (AP) on SAW and mean error on the NYUv2 data set when training our network only using real data. It is clear that without synthetic data, the performance of our network on reflectance, shading and normal shows a significant gap relative to the "full" model. This is because training a network that performs reasonably well requires a huge amount of data. The sparsity of the annotation for reflectance and shading, and the small amount of real images makes the training intractable.

Without pseudo supervision. Table 4 "w/o $\mathcal{L}_r^{ss}$" shows that on IIW and NYUv2, performance degrades relative to the "full" model, except for the AP on the SAW dataset. This shows that the self-supervision helps to provide rough guidance for the real unlabeled data on reflectance and normal. The degradation for shading is probably due to the large domain gap between the lighting of synthetic data and real data. However, when compared with shading of [23] in in Figure 6, we see even with weak supervision, our model can still predict more reasonable shading.

Contribution of multiple scales. We clearly see in Table 4 that "scale1+scale2" outperforms "scale1", and our "full" model further outperforms "scale1+scale2". It suggests that further adding a finer scale module indeed helps the local lighting modeling and boosts the overall performance. Worth noting that there is gradually saturation by further adding finer modules as the improvement gap from "scale1+scale2" to "full" is smaller than "scale1" to "scale1+scale2". In practice, we define our full model to have three scales, a coarse net with two cascaded finer nets, which strikes a good balance between accuracy and model complexity.

Without symmetric loss. The WHDR hinge loss proposed by [29] (Equation 8) is not symmetric. This leads to unequal loss when the same points are used in a different order. By adapting the WHDR to our proposed symmetric one (Equation 9), we observe improvement on IIW by 0.14%.

Model complexity: We calculate the model parameters of CGI [23] and our full model. There are $68,572,482$ floating numbers in CGI and only $14,665,594$ in our model, which is much smaller than CGI. Among the state-of-the-art CNN based methods, our method achieves consistently better performance with a smaller model size.

6. Conclusions

In this paper, we propose to estimate reflectance, normal and lighting from a single image, which is a very hard problem that has not been well addressed. A global and local SH model is proposed to model the lighting of a natural scene which accounts for both holistic lighting and the spatial variation of the lighting. A novel non-negative constraint is proposed to force the SH lighting to be physically meaningful. A synthetic data set is applied as augmentation for real data. Extensive experiments on SAW, IIW, and NYUv2 dataset demonstrate the effectiveness of our proposed method.

7. Acknowledgement

This work is supported by DARPA MediFor program under cooperative agreement FA87501620191, Physical and Semantic Integrity Measures for Media Forensics.

References

[1] Jens Ackermann and Michael Goesele. A survey of photometric stereo techniques. *Found. Trends. Comput. Graph. Vis.*, 9(3-4), 2015. 1, 2

[2] Jonathan T Barron and Jitendra Malik. Intrinsic scene properties from a single rgb-d image. In *CVPR*, 2013. 1, 3, 4, 5, 6, 7

[3] Jonathan T Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *TPAMI*, 2015. 1, 2, 4

[4] Harry G. Barrow and Jay M. Tenenbaum. Recovering intrinsic scene characteristics from images. *Computer Vision Systems*, 1978. 1

[5] Ronen Basri and David W. Jacobs. Lambertian reflectance and linear subspaces. *TPAMI*, 25(2), 2003. 1, 2, 3, 4

[6] Peter N. Belhumeur, David J. Kriegman, and Alan L. Yuille. The bas-relief ambiguity. *IJCV*, 35(1), 1999. 2

[7] Sean Bell, Kavita Bala, and Noah Snavely. Intrinsic images in the wild. In *SIGGRAPH*, 2014. 2, 3, 4, 5, 6, 8

[8] Sai Bi, Xiaoguang Han, and Yizhou Yu. An l1 image transform for edge-preserving smoothing and scene-level intrinsic decomposition. *ACM ToG*, 34(4), 2015. 3, 8

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

[10] Lechao Cheng, Chengyi Zhang, and Zicheng Liao. Intrinsic image transformation via scale space decomposition. In *CVPR*, 2018. 3

[11] Jean-Denis Durou, Maurizio Falcone, and Manuela Sagona. Numerical methods for shape-from-shading: A new survey with benchmarks. *CVIU*, 109(1), 2008. 2

[12] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *ICCV*, 2015. 3

[13] Qingnan Fan, Jiaolong Yang, Gang Hua, Baoquan Chen, and David Wipf. Revisiting deep intrinsic image decomposition. In *CVPR*, 2018. 3, 7, 8

[14] Elena Garces, Adolfo Munoz, Jorge Lopez-Moreno, and Diego Gutierrez. Intrinsic images by clustering. *Comput. Graph. Forum*, 31(4), 2012. 8

[15] Peter Vincent Gehler, Carsten Rother, Martin Kiefel, Lumin Zhang, and Bernhard Schölkopf. Recovering intrinsic images with a global sparsity prior on reflectance. In *NIPS*, 2011. 2

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

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

[18] Michael Janner, Jiajun Wu, Tejas D. Kulkarni, Ilker Yildirim, and Josh Tenenbaum. Self-supervised intrinsic image decomposition. In *NIPS*, 2017. 2

[19] Junho Jeon, Sunghyun Cho, Xin Tong, and Seungyong Lee. Intrinsic image decomposition using structure-texture separation and surface normals. In *ECCV*, 2014. 3

[20] Seungryong Kim, Kihong Park, Kwanghoon Sohn, and Stephen Lin. Unified depth prediction and intrinsic image decomposition from a single image via joint convolutional neural fields. In *ECCV*, 2016. 3

[21] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014. 5

[22] Balazs Kovacs, Sean Bell, Noah Snavely, and Kavita Bala. Shading annotations in the wild. In *CVPR*, 2017. 2, 4, 5, 6

[23] Zhengqi Li and Noah Snavely. Cgintrinsics: Better intrinsic image decomposition through physically-based rendering. In *ECCV*, 2018. 3, 4, 5, 6, 7, 8

[24] Zhengqi Li and Noah Snavely. Learning intrinsic image decomposition from watching the world. In *CVPR*, 2018. 3

[25] Wei-Chiu Ma, Hang Chu, Bolei Zhou, Raquel Urtasun, and Antonio Torralba. Single image intrinsic decomposition without a single intrinsic image. In *ECCV*, 2018. 2

[26] Takuya Narihira, Michael Maire, and Stella X. Yu. Direct intrinsics: Learning albedo-shading decomposition by convolutional regression. In *ICCV*, 2015. 3

[27] Takuya Narihira, Michael Maire, and Stella X. Yu. Learning lightness from human judgement on relative reflectance. In *CVPR*, 2015. 3

[28] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 4, 6

[29] Thomas Nestmeyer and Peter V Gehler. Reflectance adaptive filtering improves intrinsic image estimation. In *CVPR*, 2017. 3, 4, 7, 8

[30] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 4

[31] Ravi Ramamoorthi and Pat Hanrahan. On the relationship between radiance and irradiance: Determining the illumination from images of a convex lambertian object, 2001. 1, 2, 3

[32] Ashutosh Saxena, Min Sun, and Andrew Y. Ng. Make3d: Learning 3d scene structure from a single still image. *TPAMI*, 31(5), 2009. 3

[33] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D. Castillo, and David W. Jacobs. Sfsnet: Learning shape, refectance and illuminance of faces in the wild. In *CVPR*, 2018. 2

[34] Evan Shelhamer, Jonathan T. Barron, and Trevor Darrell. Scene intrinsics and depth from a single image. In *ICCV (Workshop)*, 2015. 3

[35] Jian Shi, Yue Dong, Hao Su, and Stella X. Yu. Learning non-lambertian object intrinsics across shapenet categories. In *CVPR*, 2017. 2

[36] Sameer Shirdhonkar and David W. Jacobs. Non-negative lighting and specular object recognition. In *ICCV*, 2005. 2, 3, 6

[37] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *CVPR*, 2017. 2

[38] Robert J. Woodham. Photometric method for determining surface orientation from multiple images. *Optical Engineering*, 19, 1980. 2

[39] Ruo Zhang, Ping-Sing Tsai, James E. Cryer, and Mubarak Shah. Shape-from-shading: a survey. *TPAMI*, 21(8), 1999. 2

[40] 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. In *CVPR*, 2017. 5, 6, 7

[41] Qi Zhao, Ping Tan, Qiang Dai, Li Shen, Enhua Wu, and Stephen Lin. A closed-form solution to retinex with non-local texture constraints. *TPAMI*, 34(7), 2012. 8

[42] Hao Zhou, Jin Sun, Yaser Yacoob, and David W. Jacobs. Label denoising adversarial network (ldan) for inverse lighting of faces. In *CVPR*, 2018. 2

[43] Tinghui Zhou, Philipp Krähenbühl, and Alexei A Efros. Learning data-driven reflectance priors for intrinsic image decomposition. In *ICCV*, 2015. 3, 8

[44] Daniel Zoran, Phillip Isola, Dilip Krishnan, and William T. Freeman. Learning ordinal relationships for mid-level vision. In *ICCV*, 2015. 3