CGI2018 manuscript No.
(will be inserted by the editor)

Deep Inverse Rendering for Practical Object Appearance Scan with Uncalibrated Illumination

First Author · Second Author

Abstract In this paper, we propose a practical method to estimate object appearance from an arbitrary number of images. We use a moving flashlight as light source, and capture a sequence of input photographs. Object surface material is modeled as SVBRDF properties and detail normal maps, and encoded in a pre-learned embedded latent space. Such lighting and appearance model combination enables our method to effectively narrow the solution space. On the one hand, smoothly and tightly constructed latent space allows our method to always find reasonable object appearance solutions. On the other hand, flashlight reduces the difficulty of estimating lighting conditions for each input photograph. Uncalibrated illumination requirement extremely simplifies our setup and affords it unnecessary to accurately locate light positions in advance. Moreover, our method automatically selects key frames before appearance estimation, which largely reduces calculation cost. Both synthetic and real experiments demonstrate that our method can recover object appearance accurately and conveniently.

Keywords SVBRDF · reflectance · appearance · flashlight

1 Introduction

Appearance capture is attractive but also challenging in both computer graphics and vision communities. It enables various applications in VR and AR, such as image relighting and virtual object insertion. Specially designed devices are used for accurate appearance capture [2] [6] [11]. Although these methods can reproduce high-resolution appearance, involved extensive scan effort prevents them from practical applications. In the past decade, consumer digital cameras have evolved a lot and it is quite convenient for non-expert users to capture high-quality images. For reflectance recovery, recent deep learning based methods learn shape, material priors from large-scale datasets, and take fewer images than traditional methods to inference appearance properties. Therefore, it shows good prospects to design lightweight methods based on mobile phone cameras and deep learning technologies.

Some methods [7] [18] [19] use convolutional neural networks to recover spatially variant reflectance of planar exemplars from a single image. Although they can obtain plausible results, many artifacts still exist due to insufficient observations. One approach to improve performance is to increase the number of input images. Recent methods [8] [10] take an arbitrary number of photographs to estimate SVBRDF of planar materials. Impressive results inspire us to adopt similar schemes for object surface material estimation.

In this paper, we aim to capture object appearance from multiple input photographs. We use a neural network as an optimizer to estimate SVBRDF and normal under uncalibrated flashlight illumination. Object reflectance properties are encoded in a pre-learned latent space. Such a well-constructed latent space not only promises a reasonable SVBRDF but also provides an elegant search routine towards the final solution. During optimization, we sum reconstruction loss for each input photograph together as [10] [15], and this provides flexibility about the number of input images. Experiments demonstrate that our method can recover

F. Author
first adress

S. Author
second address

SVBRDF from plausible to accurate with the increment of input image number.

For multiple images capture, one consideration is how to take photographs efficiently. Since shooting videos of objects under a moving flashlight is quite simple, we select frames from videos as input. Generally there is a trade-off between the image number and recovery accuracy. Given the budget of the input image number, we propose to select the most valuable image collection via classic clustering. Experiments show that our selection strategy chooses reasonable images collection and is favorable to recover appearance efficiently.

In summary:

- We propose a practical framework to estimate SVBRDF and normal for objects with only off the shelf devices.
- We adopt planar material latent space for object surface via normal decomposition.
- We apply key frame selection strategy to promote algorithm efficiency.

2 Related work

In graphics, forward rendering composes shape, reflectance and illumination together according to the rendering functions. Estimating object shape, reflectance properties and lighting from images is an inverse rendering problem, which has been studied for a long time.

Intensive measurement One straightforward approach to capture appearance is brute-force measurement. Researchers design professional devices to control lighting and camera views for such purpose [23] [35]. Dana *et al.* [6] used a robot arm to densely sample incident light and view combinations for planar material samples. Another kind of common devices are light stages [2], they are mounted with a large number of lights and able to provide incident light from considerable directions. Linear light source reflectometry [11] [3] is also broadly adopted for capturing spatially variant materials. Although those methods can recover vivid appearance, their dedicated devices hinder them from consumer applications.

Simplified acquisition In order to reduce the operating threshold for average users and simplify the acquisition process, some searchers try to capture appearance with hand-held commodity devices. Without precisely controlled light stages, many methods take natural environment illumination or mobile flashlights. Such as Riviere *et al.* [30] took a mobile device to record reflectance from specular material samples under paired flashlights. Other methods [29] [12] use simple tools and apply sparse priors to recover both normal and SVBRDF for planar surface.

Comparing with planar material, object appearance recovery needs to care about shape geometry. Wu *et al.* [36] took Kinect sensors to scan object geometry and acquired illumination via a mirror sphere, then computed object appearance in an inverse rendering framework. With known geometry but unknown natural illumination, Dong *et al.* [9] estimated isotropic surface SVBRDF from a video of a rotating subject. Some methods jointly solve shape and materials with single or several images as input [27]. Barron *et al.* [1] inferred priors about shape, reflectance and illumination from natural images statistics, and recovered most likely shape, reflectance and illumination from a single input image. Comparing with these methods, we take advantages of neural networks to regularize SVBRDF in a reasonable space rather than relying on hand-crafted priors or specified heuristics.

Deep inverse rendering Li *et al.* [18] proposed a novel self-augment training scheme which effectively expanded training dataset. Deschaintre *et al.* [7] utilized in-network render layers to construct reconstruction loss and estimated reflectance properties from a flash-lit single image. Similar to [7], Li *et al.* [19] benefited from in-network render during training but added a dense CRF model to refine final results. However, these methods take only one image as input, they often fail when visible reflectance features are insufficient to distinguish ambiguities.

Gao *et al.* [10] presented a novel framework taking an arbitrary number of images as input. Impressively, they optimized SVBRDF in a specially learned latent space, promoting their network to obtain reasonable estimation. Sharing the same idea to utilize multiple images, Deschaintre *et al.* adopted an order-independent fusion layer to combine multiple uncalibrated flash-lit images feature together. These methods focus on near-planar material sample, in contrast, ours method focus on object appearance recovery. Recently, Li *et al.* [20] proposed a learning-based method to jointly regress shape, SVBRDF and illumination from a single flash-lit image. Like previous single input image methods, it suffers from insufficient observations.

Another related work is [14], they designed an asymmetric deep auto-encoder to model image formation and inverse rendering process. Trained with a large amount of synthetic data, the auto-encoder automatically learned optimal lighting patterns. Extended from [14], Kang *et al.* [15] utilized learned lighting pattern to efficiently capture object appearance, and exploited diffuse and

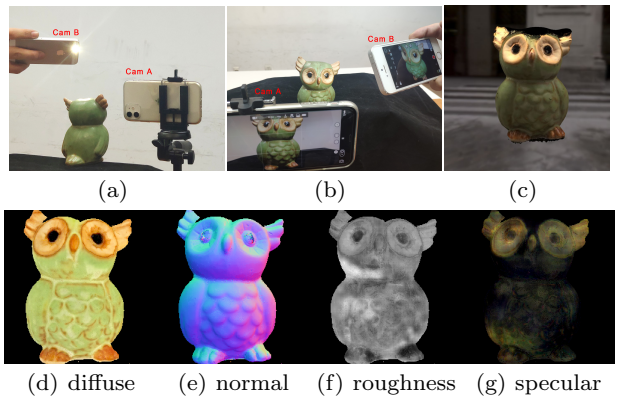


Fig. 1 (a) - (b): Capture setup. During capture, all lights in the room are turned off. Cam_A is fixed to shoot videos of target object. Cam_B serves as a light source.(c) Reconstructed object. (d) - (g): estimated SVBRDF and normal.

normal information from multiple views to reconstruct geometry.

Uncalibrated photometric stereo Our framework shares similar inputs with uncalibrated photometric stereo methods. We both estimate object normal from multiple images under different unknown illumination. Classic photometric stereo methods explore effective cues from color and intensity priors [32] [22], reflectance symmetry [37] [21]. Recently, learning based methods DPSN [31], PS-FCN [5] and CNN-PS [13] show promising normal estimation results. Chen *et al.* [4] proposed a two stage method: they first regressed light directions and intensities in LCN, then estimated final normal. A comprehensive evaluation of uncalibrate photometric stereo can be found in [33].

3 Method

3.1 Preliminary

Our goal is to estimate object SVBRDF and normal in a single view. As showed in Fig 1, a fixed camera is deployed to capture object-center images while a flashlight is moving. The power distribution of the flashlight is roughly concentrated in a solid angle. Therefore, we model the flashlight as a point light source as long as keep it facing the target object during the capture process. We assume that camera inner parameters are fixed, field of view fov is known, and flashlight intensity keeps constant as I_{int}. We adopt the Cook-Torrance microfacet BRDF model with the GGX normal distribution [34] and assign BRDF parameters for each point p: diffuse ableo $k_d(p)$, specular albedo$k_s(p)$ and monocular roughness $\alpha(p)$. In conclusion, our method solves

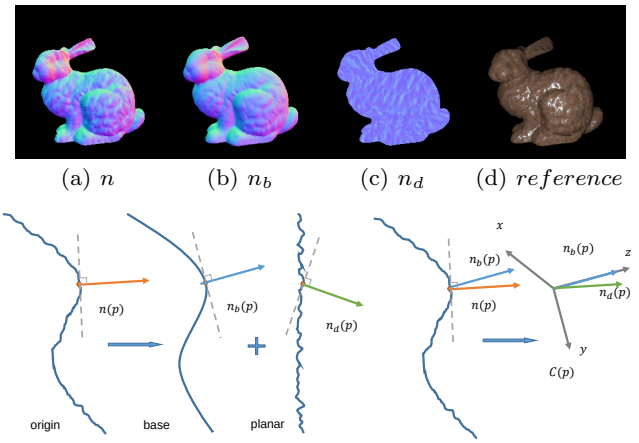


Fig. 2 Decouple shape and surface material. First row: we show normal maps including normal n, base normal n_b and detail normal n_d of $reference$. Second row: We show 2D slice of object surface. Origin object can be decomposed as base shape and planar material samples.

SVBRDF and normal $n(p)$ for target object in a fix view, with flashlight position l_i unknown for each input photograph I_i.

3.2 Decouple shape and surface material

Demonstrated in [10], a pre-learned latent space can effectively model appearance of planar exemplars and benefits SVBRDF estimation. In our case, we argue that object surface material can be modeled like planar material samples too. It is workable to use the expressive latent space in object SVBRDF recovery.

As showed in Fig 2, object surface can be viewed as warped planar material samples. We decompose normal $n(p)$ as base normal $n_b(p)$ and detail normal $n_d(p)$. Base normal $n_b(p)$ relates with object shape, and detail normal $n_d(p)$ reflects material characteristics. For each point, a local coordinate system $C(p)$ can be constructed: base normal $n_b(p)$ direction is assigned as local z axis $z(p)$ and local y axis is assigned to be perpendicular to direction $(1, 0, 0)$. In such a local space, detail normal $n_d(p)$ means deviation from base normal $n_b(p)$ and engraves detail variation. Therefore we get:

$$n(p) = n_b(p) \circ n_d(p) \tag{1}$$

Operation $\circ$ means transforming $n_d(p)$ from local space $C(p)$ (constructed according to $n_b(p)$) into global space. It crosses the gap between complex shape and planar material samples. Then object surface materials can be naturally modeled in the planar material latent space.

Once shape and surface material are decoupled, it is easy to convert lighting and viewing directions into each point' local coordinates. Solving object SVBRDF

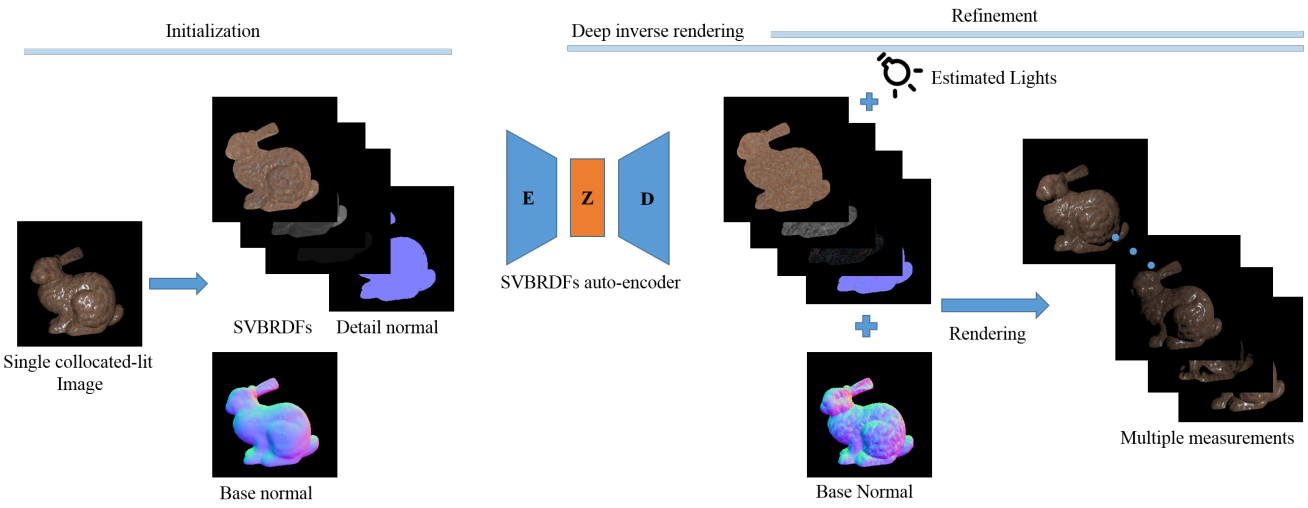


Fig. 3 Overview of the deep inverse rendering framework. We use method [20] to estimate SVBRDF and base normal from single collocated-lit image as initialization. A specially designed SVBRDF auto-encoder is adopted for deep inverse rendering: we adjust z code to decode SVBRDF and update simultaneously base normal and point light positions. Finally, we directly refine all components in the post-processing step.

and detail normal $n_d(p)$ in local coordinate is equal with planar material samples appearance recovery. Because we adopt local lighting model, object position is needed to calculate light direction and intensity attenuation for each point. But, in fact, the distance d_i between the flashlight and object obj is much larger than the scale of object geometry variation. Thus, with known fov, a rough depth map in the camera view is enough to calculate object position. Li *et al.* [20] estimated depth, normal and SVBRDF from a single image lit by a flashlight, which provides good SVBRDF initialization and depth inference for our method.

3.3 Reflectance recovery under uncalibrated illumination

Object shape, material, and illumination jointly decide how the object looks like. If illumination is under control, SVBRDF estimation would be easier. In this paper, we drop fully control of illumination and each input photograph is lit by a flashlight with unknown position l_i. In addition, we assume flashlight intensity I_{int} is constant during the whole capture process.

As showed in Fig 3, our method consists of three stages: initialization, deep inverse rendering and refinement. The core of our method is deep inverse rendering stage. We constrain object reflectance characteristics in a pre-trained latent space [10] and adjust the latent code z to decode reflectance parameters s:

$$s = D(z), \tag{2}$$

where $s = (n_d, k_d, \alpha, k_s)$.

We formulate the deep inverse rendering as a minimization that jointly updates the latent code z, base normal n_b and light positions $\{l_i\}$ to minimize the differences between input photograph I_i and corresponding rendering image $R(s, n_b, l_i)$:

$$\arg\min_{z, n_b, \{l_i\}} \sum_i \mathcal{L}(I_i, R(D(z), n_b, l_i)). \tag{3}$$

where we use the common loss function [7] [10] as:

$$\mathcal{L}(x, y) = ||log(x + 0.01) - log(y + 0.01)||_1. \tag{4}$$

The whole pipeline of our method is as follows:

1. Use existing methods to initialize base normal, depth and SVBRDF with single collocated-lit image I_{col} as input.
2. Search initialized light positions for each input photograph I_i.
3. Optimize latent code z , base normal n_b and light positions $\{l_i\}$ in deep inverse rendering stage.
4. Image space refinement for SVBRDF, detail normal n_d, base normal n_b and light positions $\{l_i\}$.

Reflectance initialization We capture a special image I_{col} whose correspondent flashlight is collocated with camera lens, as previous methods [20] [24] [19]. I_{col} (we called collocated-lit image) has following advantages: 1. It has almost no shadow; 2. During optimization, light position l_{col} is fixed and known. Besides, I_{col} can be used [20] to initialize our reflectance properties map. The method takes I_{col} as input, and estimates object depth $\hat{d}$, SVBRDF $\hat{s} = (k_d, \alpha, k_s)$ and normal $\hat{n}$ in a

cascaded network. We take their estimated normal $\hat{n}$ as base normal n_b and initialize n_d as a flatten normal map ($n_d(p) = (0, 0, 1)$). In addition, given camera fov, we project depth map $\hat{d}$ into 3D coordinates as object position map Pos.

Light initialization Initialized SVBRDF $\hat{s}$ and normal $\hat{n}$ would help us initialize light position $\hat{l}_i$ for each photograph I_i. We use a try-and-compare strategy to search for light position candidates.

First, we define the metric to evaluate the possibility that a light position candidate lc_i can be used to initialize $\hat{l}_i$. For input image I_i, we use $\hat{s}$, $\hat{n}$ and lc_i to render image R_i, and calculate RMSE for R_i. Therefore, the goal of light initialization is to quickly find lc_i who has smaller RMSE for R_i.

Then we search light position candidate $\acute{l}c_i$ in iterations. 1) At first, we construct a rectangular cuboid centered on the camera. During image capture, the flashlight is moving around the camera. Therefore, we set the cuboid size as $height = w_h * \bar{d}, length = w_d * \bar{d}, depth = w_d * \bar{d}$, where $\bar{d}$ is the average of depth map $\hat{d}$. We draw grids with step (w_s, l_s, d_s) in the cuboid and find the current best $\acute{l}c_i$ from all vertices. 2) Next we construct downsized cuboid centered on $\acute{l}c_i$, and draw downsized grids in the new cuboid. Similarly, we find new $\acute{l}c_i$ from all vertices. We iterate the search process until $\acute{l}c_i$ is not updated or cuboid size is below the threshold.

Finally, for each photograph I_i, $\acute{l}c_i$ is used as initialized light position $\hat{l}_i$.

Image space refinement In our optimization network, latent code encodes reflectance properties in the bottle neck of auto-encoder. It usually decodes reflectance properties with details lost. Thus, we add post-process step to refine SVBRDF properties maps pixel by pixel [10]. Instead of adjusting the latent code z, we directly update SVBRDF parameters k_d,α,k_s, detail normal n_d and light positions $\{l_i\}$ to minimize the differences between I_i and rendering image $R(s, n_b, l_i)$. We formulate the image space refinement as:

$$\underset{k_d,\alpha,k_s,n_d,n_b,\{l_i\}}{\arg\min} \sum_i \mathcal{L}(I_i, R(s, n_b, l_i)). \tag{5}$$

3.4 Key frames selection

More input images captured under different illumination help dissolve ambiguity and achieve accurate reflectance recovery. But it is not feasible to take hundreds of images as input due to expensive calculation effort. Moreover, redundant image input may contribute little and it is reasonable to discard those invalid ones.

Optimal BRDF sampling for near-field reflectance acquisition has been discussed in previous works [26] [38]. Similarly, object SVBRDF estimation has optimal light and view direction pairs. When the camera view is fixed, object shape and materials jointly determine optimal light directions. However, we do not control light position explicitly. It is not feasible for us to adopt optimal lighting strategy like previous methods. Therefore, we decide to directly select images where the target object shows the distinctive appearance. When photographs look similar, they are possibly lit by flashlights close to each other. Thus, selecting different looking images means choosing different lighting directions. Here we rely on the classic k-means clustering method to divide all captured images into different clusters, and select centroids as picked images. Given recorded videos, our strategy can free users from tedious manual image selection and promotes efficiency.

3.5 Multiple views

SVBRDF capture in a fixed single view can be considered a sub-task of recovering an object's appearance from numerous directions. Our method can be extended into multiple views, and reconstruct 3D shape with estimated SVBRDF texture. Following the geometry reconstruction [15], we take diffuse albedos and normal maps as input for multiple view stereo. Estimated normal maps contain abundant geometric information, which is beneficial to surface detail recovery. Thus we refine reconstructed shape with estimated normal maps to obtain the final geometry. Finally, we merge estimated SVBRDF from different views into the texture space.

Our reconstruction pipeline is as follows:

1. Estimate SVBRDF and normal n_i independently for each view.
2. Run sparse reconstruction with diffuse albedos and obtain camera external parameters for each view.
3. Transfer normal from camera space into world space.
4. Run dense reconstruction with diffuse albedos and transferred normal maps $\{n_i'\}$ respectively to get two point clouds.
5. Merge two point clouds and run Poisson reconstruction [16] to get a rough geometry.
6. Project each vertex of the geometry into each view. If a vertex v is visible in I_i, record corresponding normal $n_i(p)'$ and calculate the angle between camera view and $n_i(p)'$ as weights. Then update vertex normal n_v with weighted average of $n_i(p)'$.
7. Refine the geometry with updated normal [25].
8. Generate UV maps using UVAtlas [40] and merge SVBRDF estimation from visible views.

4 Results

We implemented our method in Tensorflow and take built-in layers to construct a differentiable render. For SVBRDF auto-encoder, we inherit trained model from [10]. We choose Adam [17] as optimizer, setting learning rate as 10^{-3} and β_1 as 0.5. In deep inverse rendering stage, we run $6k$ iterations; In refinement stage, we run $1k$ iterations.

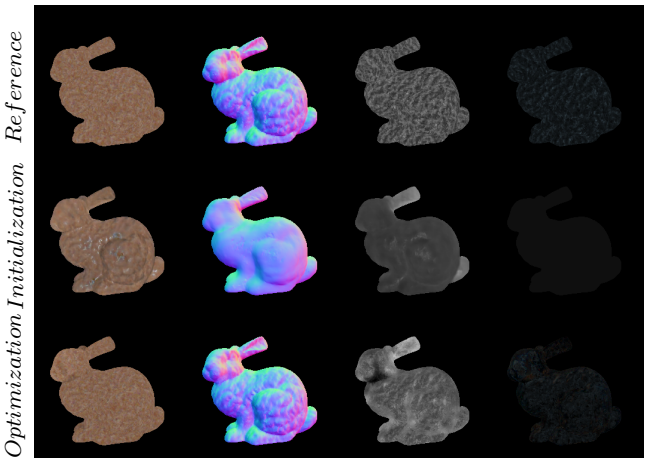
At the beginning, we create synthetic datasets to validate proposed method. We randomly compose distorted elementary shapes into synthetic objects like [39] and apply texture from materials dataset [7]. In addition to composed shapes, we also select several models from the Standford 3D Scanning Repository. We render images with pre-defined point lights (used to approximate real flashlights) in a rectangle area.

To demonstrate the effectiveness of the whole method, we gradually relax the restriction from known lighting positions to unknown.

4.1 Known lighting

First, we validate whether pre-trained planar materials appearance latent space can be used in object SVBRDF estimation. For the synthetic experiments, we set camera fov as 60 degree, image resolution as 256x256, and the distance between the camera and objects as $2units$. Suppose the camera center is C and the target object is at point O. All point lights are located insides the plane which is perpendicular to the line CO. In the 2x2 $unit^2$ rectangular area, we uniformly place point lights at vertices of the 5x5 grid. These point lights are used to render LDR input images I_i. For testing, we sample point light positions in the 4x4 grid, crossly among 5x5 grid. Given ground truth light positions l_i for each input photograph I_i, we take multiple images as inputs to estimate appearance properties. To quantitatively evaluate our methods, we adopt metrics as follows: 1) RMSE(root mean square error) for estimated diffuse, specular and roughness albedos against ground truths. $s_{est} = (k_s, k_d, \alpha)$; 2) normal deviation between estimated normal n_{est} and ground truth n_{gt} in degree; 3) RMSE for rendering images $R'(n_{est}, s_{est}, \tilde{l}_i)$ under test lightings $\{\tilde{l}_i\}$. Note that SVBRDF albedos and rendering images are normalized in [0,1] to calculate RMSE.

We test on 21 objects with different materials and show average error in the Table 1. All SVBRDF properties errors are smaller than initialization, and normal accuracy has been improved impressively. In general, our results are much closer to the reference than initialization. We show results of bunny in Fig 4. Comparing with initialization, less highlight artifacts show in estimated diffuse albedo, and our estimated normal map contains more details.



(a) Diffuse (b) Normal (c) Roughness (d) Specular

Fig. 4 Bunny results with known lighting.

Table 1 Results with known lighting. Init: initialization; Opt: optimization.

error	diffuse	roughness	specular	normal	render
Init	0.0811	0.1652	0.1096	19.47	0.0976
Opt	0.0189	0.0729	0.0706	1.2235	0.0548

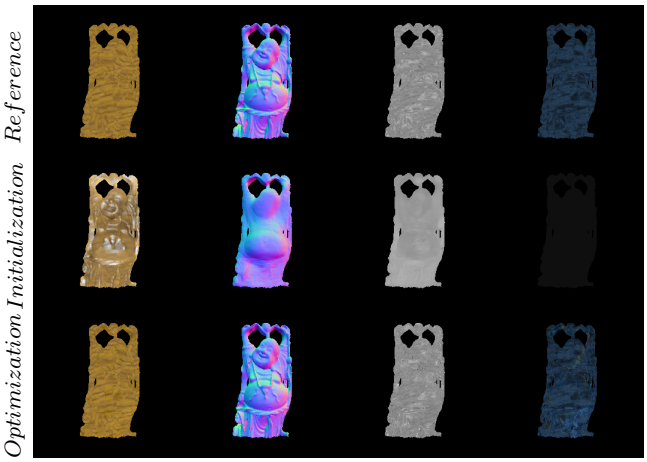
Table 2 Results with unknown lighting. Init: initialization; Opt: optimization. Light: light position.

error	diffuse	roughness	specular	normal	render	light
Init	0.0811	0.1652	0.1096	19.47	0.0976	-
Opt	0.0200	0.0716	0.0706	1.3001	0.0549	0.0398

4.2 Reflectance recovery under uncalibrated illumination

Uniformly sampled lights At first, we takes 25 images lit by uniformly sampled lights as input, and still test on synthetic objects mentioned in Sec 4.1 (Without special statement, we take such 21 objects for synthetic experiments by default). Light positions estimation will be measured in distance. We show average error in Table 2. Comparing with initialization, all refectance properties have been improved, and RMSE for render images is lower. At the same time, estimated light positions are close to the actual light positions. We show optimization result of buddha in Fig 5 and light position estimation in Fig 6. After optimization, light positions converge to ground truth dramatically.

K-means frames selection In synthetic experiments, we first render images with 400 uniformly sampled lights from a grid of 20x20 in the same rectangular area as



(a) Diffuse (b) Normal (c) Roughness (d) Specualr

Fig. 5 Buddha result with unknown lighting.

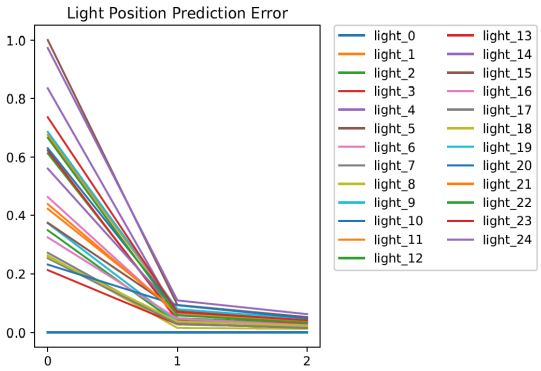


Fig. 6 Light estimation error for buddha. Vertical axis records distance between predicted light position and ground truth. Horizontal axis indicates three stages: initialization, deep inverse rendering and refinement.

mentioned in Sec 4.1. Besides uniformly sampled lights, we render collocated-lit image I_{col} that flashlight shares the position with camera. We resize those images as vectors and run k-means algorithm to cluster them into several groups. After picking centroids from all groups, we find the group which collocated-lit image I_{col} belongs to and replace its centroid by I_{col}. Therefore, we always keep I_{col} in our final selected image collections.

In Fig 7, we show k-means clustering results for bunny. Each point represents a image lit by a flashlight. Since all lights in synthetic experiments are sampled in a rectangular planar, we take (x, y) from actual light position (x, y, z) as 2D coordinates to draw points in the figure. We observe that k-means clustering results are coincident with flashlight positions. If some flashlights are close, their correspondent images will be clustered in the same group. We also show selected images in Fig 8: each picked image has different highlight areas and highlights uniformly cover the whole body of bunny,

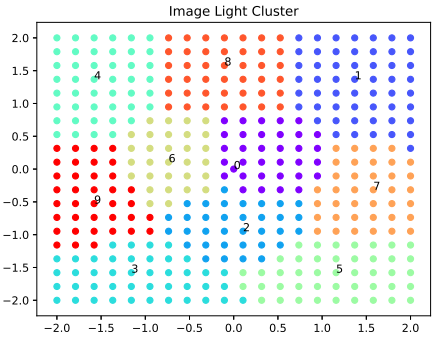


Fig. 7 Bunny k-means clustering result. 401 images are clustered into 10 groups. Images in each group share the same color and centroids are labeled with cluster id.

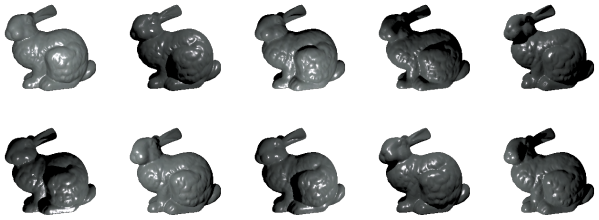


Fig. 8 K-means clustering results for bunny. 10 distinctive images are selected from bunny image collection.

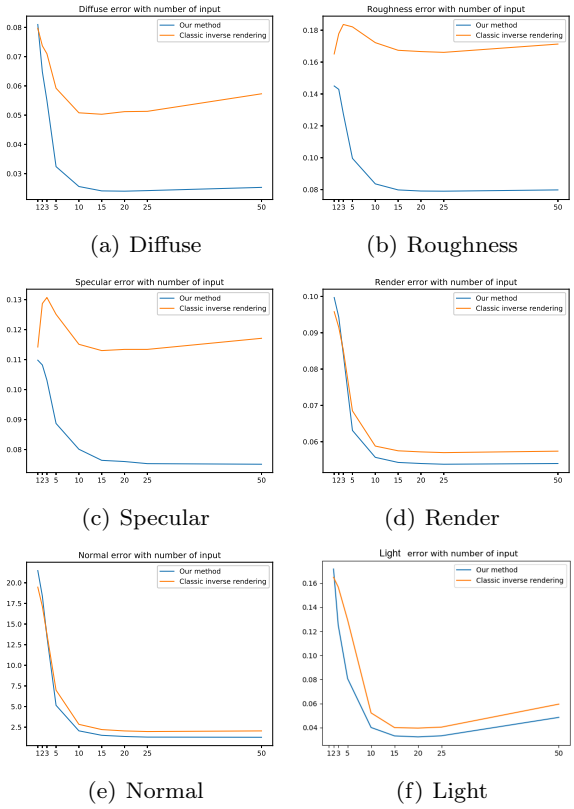
providing cues to estimate SVBRDF and normal. It illustrates that our k-means selection strategy automatically selects distinctive images and ensures lighting variation.

Next, we select 25 images and summarize optimization results in Table 3. All SVBRDF properties, render images quality, and normal estimation have been improved significantly. Comparing with optimization results of 25 uniformly sampled flashlights in Table 2, SVBRDF properties and normal are slightly worse but RMSE of render images is every close. That is because uniformly sampled 25 images contain flashlights on the border of grid and cover larger incident lighting direction scope. Since the difference in optimization result is slight and both are dramatically better than initialization, we argue that our images selection strategy can automatically select valuable images. As showed in Fig 14, initialization method [20] cannot distinguish diffuse and specular components clearly. Their method misses normal details, leaving variations in diffuse albedo. In comparison, our method takes multiple images to estimate accurate SVBRDF, and rendering images under novel lighting look almost similar with ground truth.

Number of input images In this section, we show how the number of input images affects optimization results. We use k-means strategy to select images with the num-

Table 3 Result with k-means selected 25 images under unknown lighting. Init: initialization; Opt: optimization.

error	diffuse	roughness	specular	normal	render	light
Init	0.0811	0.1652	0.1096	19.47	0.0976	-
Opt	0.0267	0.1048	0.0732	2.1573	0.0545	0.0448

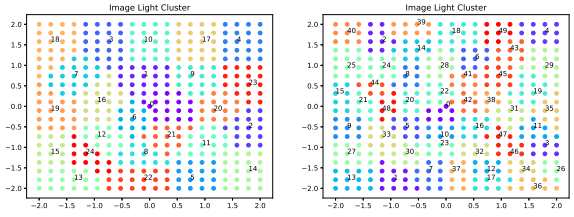


(a) Diffuse (b) Roughness

(c) Specular (d) Render

(e) Normal (f) Light

Fig. 9 Optimization result with different number of input.



(a) K-means 25 (b) K-means 50

Fig. 10 K-means clustering comparison.

ber k ranging from 1 to 50 and show comparison in Fig 9. In general, as the increment of input image number k, diffuse, specular and roughness estimation performance improves. Estimated normal becomes dramatically accurate with more input images. When k comes to 10, the rate of improvement slows. Another key point is 25, more images than 25 bring little benefits.

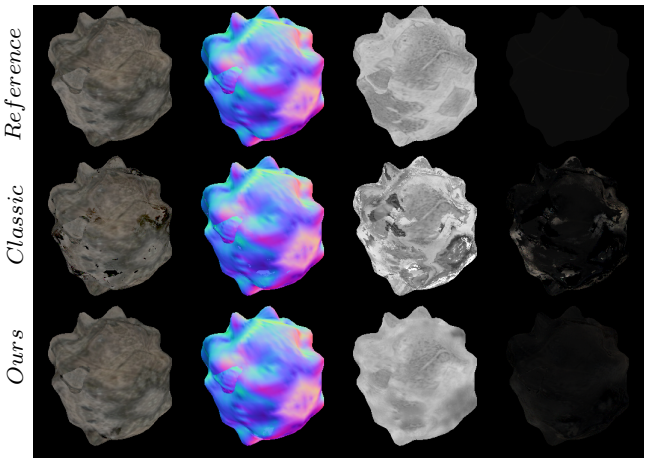
One exception is about light estimation. When only collocated-lit image I_{col} is fed into our network, average light position estimation error is zero. Then more input images promote both light and SVBRDF estimation. We argue that accurate light and SVBRDF estimation are favorable to each other, and they converge towards ground truth together. However, when k is beyond 25, light estimation error enlarges. This is because when 50 input images are selected, images illuminated by the flashlights who are far from the camera center will be selected. The distance between those flashlights and the target object is so great that a small deviation from the ground truth position will only cause minor variation in the direction of incident light. Although absolute light position estimation error rises, other SVBRDF properties and normal estimation still keep accurate. Comparison between selected 25 and 50 images distribution is showed in Fig 10. It indicates that when 50 images are selected, more images are illuminated by flashlights around the border.

Comparison with classic inverse rendering To prove optimization in the latent space is actually helpful, we compare our method with classic inverse rendering that directly optimizes SVBRDF and normal in image space. Similar with the post-process step introduced in Sec 3.3, we take the differentiable render in our network to implement classic inverse rendering. Instead of updating both base normal n_b and detail n_d, classic inverse rendering directly adjust global normal n. We formulate the classic inverse rendering as:

$$\arg\min_{k_d,\alpha,k_s,n,\{l_i\}} \sum_i \mathcal{L}(I_i, R(k_s, k_d, \alpha, n, l_i)). \tag{6}$$

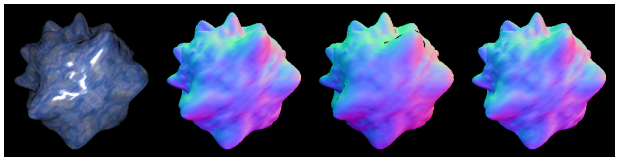
We provide the classic inverse rendering with the same initialization and run sufficient number of iterations to make sure convergence. Fig 9 shows that our method recovers more accurate SVBRDF and normal. As showed in Fig 11, classic inverse rendering optimization results contain artifacts that diffuse components are recognized as rough highlight. Although classic rendering recovers plausible normal, it fails to distinguish diffuse and specular reflection.

Comparison with uncalibrated photometric stereo We evaluate our method on real-world benchmark dataset DiLiGenT [33]. DiLiGenT provides 96 real-world images for each object, and those images are shot under different LED lights. First we choose the image whose corresponding incident light direction is closest with camera view as collocated-lit image I_{col}, then we select 10, 25 images following our strategy. We collect evaluation from the state-of-art method [4] in Table 4:



(a) Diffuse (b) Normal (c) Roughness (d) Specualr

Fig. 11 Comparison between classic inverse rendering and ours.



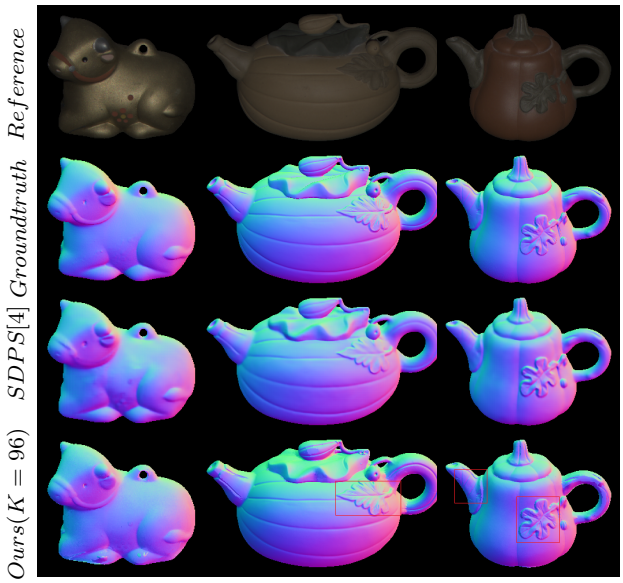
(a) Reference (b) GT (c) SDPS-Net (d) ours

Fig. 12 Normal estimation result on our synthetic datasets.

our method bypasses most methods except [4]. Comparing with SDPS-Net, our method achieves comparable performance apart from harvest and reading cases. As showed in Fig 13, our method is able to keep rich details. In addition, our method recovers both SVBRDF and normal, but SDPS-Net [4] mainly focuses on normal estimation. We also test SDPS-Net [4] with 25 input images on our synthetic datasets, average normal error is 15.7 degree, worse than our method whose error is 2.2 degree. Since our synthetic object surface contains abundant material variation and reveals complex appearance, it is difficult for [4] as showed in Fig 12.

4.3 Real Acquisition Results

We extend our method to real acquisition data. Fig 1 elucidates how devices are deployed in a dark room. We use the ProCam app in iPhone 11 to capture all images and videos. Before image capture, we manually adjust ISO, white-balance, aperture and shutter speed parameters. During image capture, all camera configuration are fixed. For each object, we first shot the collocated-lit image, then turn video mode on while moving Cam_B manually around the target. Since we have no ground truth SVBRDF and normal maps like synthetic experiments, we evaluate our method by comparing shot images and rendering images under novel lighting. For testing, we shot some pictures under novel lighting and locate their light positions. Specifically, we place a checker around the target, Cam_B shots images of the checker to locate itself. Then we take Cam_B location to approximate its flashlight position so that we obtain images with known lightings as reference.



(a) Cow (b) Harvest (c) Pot2

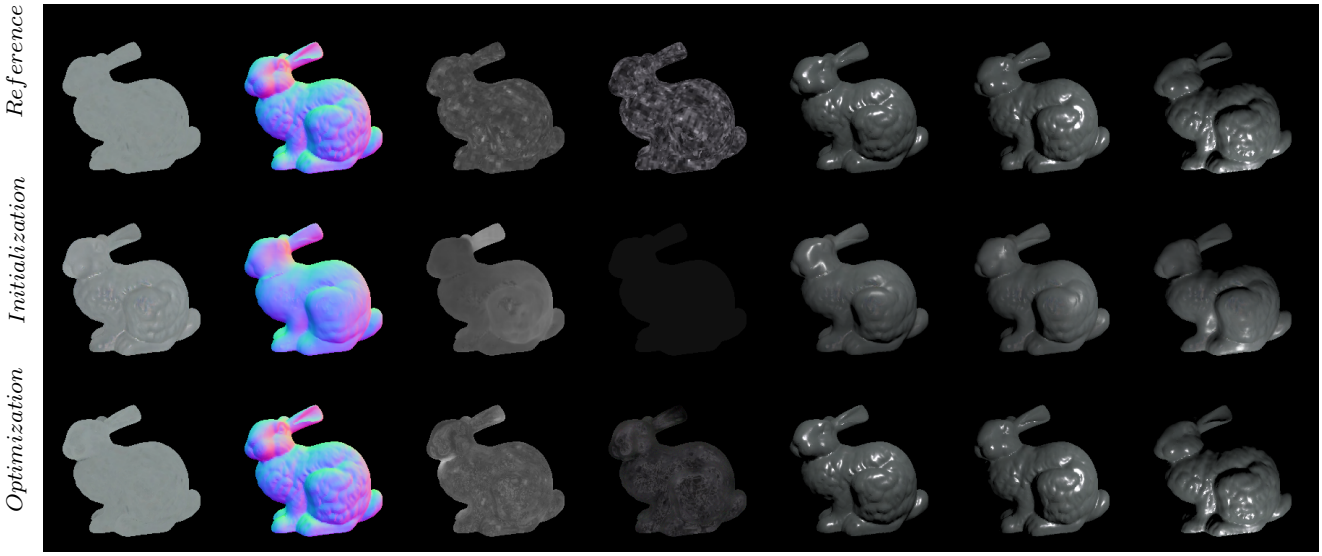
Fig. 13 Normal estimation result on DiLiGenT. Our method keeps more rich details like in red boxes.

In Fig 15, we show an example of real acquisition. Compared with initialized SVBRDF and normal, our method produces high-quality diffuse and normal map with more details. For roughness and specular map, our method may produce larger roughness estimation when highlight cues are missing, especially in the boarder area where normal is almost perpendicular to view direction. It is natural that when highlights are not observed, accurate spatial variant roughness estimation is very difficult for our optimization based framework.

We take novel lighting to render images, and compare them with references. Reference images look brighter than input images. That is because in practice reference images are shot with checker around target object, inter-reflection happens. Since the flashlight is still dominant, we believe that the more render images look like references, the better SVBRDF and normal estimation model object appearance. Our render results are very close to references: highlight appears correctly and image intensity distribution is visually consistent with references. Realistic rendering images illustrate that our method can recover object appearance effectively in real scenario.

Table 4 Normal estimation for the DiLiGenT benchmark. We select K input images for each object.

method	ball	cat	pot1	bear	pot2	buddha	goblet	reading	cow	harvest	avg.
SM10 [32]	8.90	19.84	16.68	11.98	50.68	15.54	48.79	26.93	22.73	73.86	29.59
WT13 [37]	4.39	36.55	9.39	6.42	14.52	13.19	20.57	58.96	19.75	55.51	23.93
LM13 [22]	22.43	25.01	32.82	15.44	20.57	25.76	29.16	48.16	22.53	34.45	27.63
PF14 [28]	4.77	9.54	9.51	9.07	15.90	14.92	29.93	24.18	19.53	29.21	16.66
LC18 [21]	9.30	12.60	12.40	10.90	15.70	19.00	18.30	22.30	15.00	28.00	16.30
UPS-FCN [5]	6.62	14.68	13.98	11.23	14.19	15.87	20.72	23.26	11.91	27.79	16.02
SDPS-Net [4](K=10)	3.27	10.48	9.89	6.60	9.01	10.71	11.54	17.89	10.21	19.22	10.89
SDPS-Net [4](K=25)	3.64	8.73	8.41	6.98	7.87	9.32	12.03	16.16	8.55	18.13	9.96
SDPS-Net [4](K=96)	**2.77**	**8.06**	**8.14**	6.89	7.50	**8.97**	11.91	**14.90**	8.48	**17.43**	**9.51**
Ours(K=10)	3.83	9.20	11.21	5.75	7.76	13.65	11.11	27.04	8.52	25.91	12.40
Ours(K=25)	3.50	9.03	11.89	4.69	7.36	12.41	10.18	27.65	7.10	28.46	12.23
Ours(K=96)	3.44	9.10	11.98	**4.83**	**7.28**	12.10	**10.00**	26.45	**6.34**	25.41	11.69



(a) Diffuse (b) Normal (c) Roughness (d) Specular (e) Render (f) Render (g) Render

Fig. 14 Results for bunny with 25 selected images. Each row shows SVBRDF properties, normal map and 3 rendering images under novel lighting.

4.4 Multiple Views

We take widely used Colmap for both sparse and dense reconstruction. Since 256x256 resolution images contain insufficient number of feature points for sparse reconstruction, we change image resolution into 512x512. In synthetic experiments, we choose 36 views horizontally surrounding target objects, In real acquisition, we still capture images in 36 views but keep camera fixed and rotate object. Fig 16 shows rendering images for reconstructed objects.

5 Conclusions and Future Work

We propose a lightweight method to recover object reflectance with uncalibrated flash lighting. Modeling object surface material in a pre-learned latent space enables our method to always recover reasonable SVBRDF and constrain optimization routine to reduce ambiguity. Key frames selection strategy reduces both capture and calculation cost. Moreover, our method can be extended from single view to multiple view stereos. Synthetic and real experiments show that our method can recover accurate SVBRDF and normal efficiently.

One limitation of our method is that we ignore interreflection among object components. Our method may fail when object contains large parts of concave areas. Therefore, in the future we will add multiple bounce reflection estimation modules. Currently, our method may fail if other indoor lights are not switched off. That is because when the flashlight is moving, other uncontrolled lights cast varying shadows on the target object. Thus, we consider add accurate shadow detection modules in the future.

6 Acknowledgments

The authors would also like to thank the anonymous referees for their valuable comments and helpful suggestions.

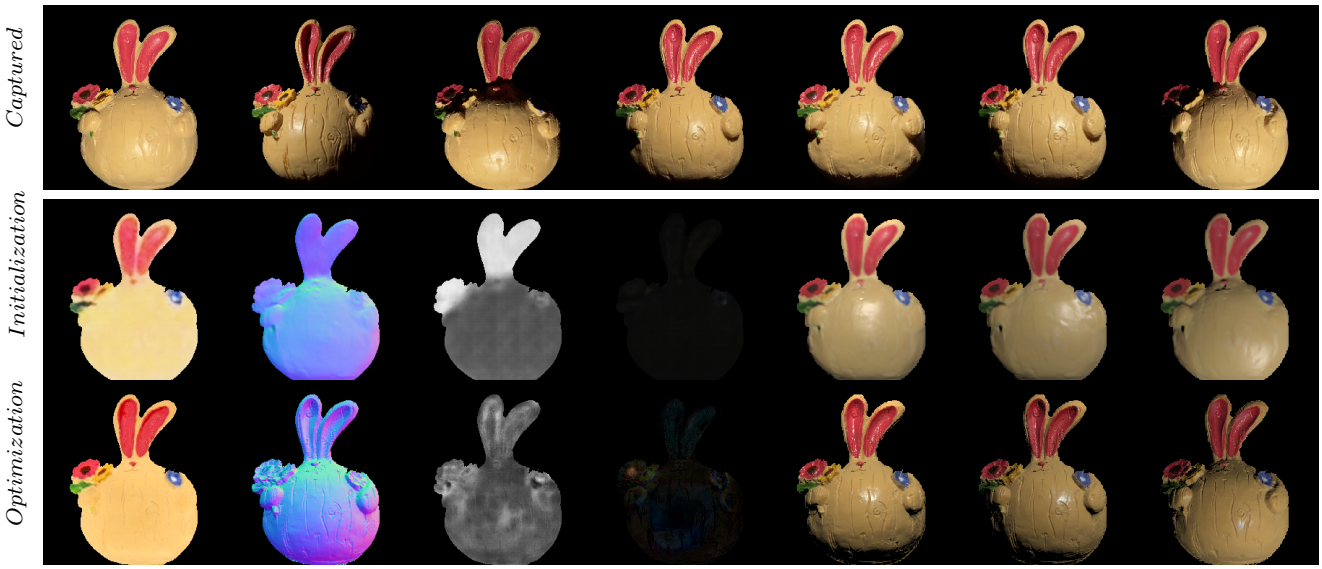


Fig. 15 Result for real captured object with 25 selected images. First row shows some captured images: left four are part of input, and another three are reference images. In second and third rows: we display SVBRDF properties, normal map and 3 rendering images under novel lighting. Each render image column shares the same novel lighting.

Fig. 16 Reconstructed object visualization: synthetic piggy and real owl.

References

1. Barron, J.T., Malik, J.: Shape, illumination, and reflectance from shading. IEEE Transactions on Pattern Analysis and Machine Intelligence **37**(8), 1670–1687 (2015)
2. Ben-Ezra, M., Wang, J., Wilburn, B., Li, X., Ma, L.: An LED-only BRDF measurement device. 26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR (i) (2008)
3. Chen, G., Dong, Y., Peers, P., Zhang, J., Tong, X.: Reflectance scanning: Estimating shading frame and BRDF with generalized linear light sources. ACM Transactions on Graphics **33**(4), 1–11 (2014)
4. Chen, G., Han, K., Shi, B., Matsushita, Y., Wong, K.K.K.: Self-calibrating deep photometric stereo networks. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8731–8739 (2019)
5. Chen, G., Han, K., Wong, K.Y.K.: Ps-fcn: A flexible learning framework for photometric stereo. In: The European Conference on Computer Vision (ECCV) (2018)
6. Dana, K.J., Van Ginneken, B., Nayar, S.K., Koenderink, J.J.: Reflectance and texture of real-world surfaces. ACM Transactions on Graphics **18**(1), 1–34 (1999)
7. Deschaintre, V., Aittala, M., Durand, F., Drettakis, G., Bousseau, A.: Single-image svbrdf capture with a rendering-aware deep network. ACM Transactions on Graphics (SIGGRAPH Conference Proceedings) **37**(128), 15 (2018)
8. Deschaintre, V., Aittala, M., Durand, F., Drettakis, G., Bousseau, A.: Flexible svbrdf capture with a multi-image deep network. Computer Graphics Forum (Proceedings of the Eurographics Symposium on Rendering) **38**(4) (2019)
9. Dong, Y., Chen, G., Peers, P., Zhang, J., Tong, X.: Appearance-from-motion: Recovering spatially varying surface reflectance under unknown lighting. ACM Transactions on Graphics **33**(6) (2014)
10. Gao, D., Li, X., Dong, Y., Peers, P., Xu, K., Tong, X.: Deep inverse rendering for high-resolution svbrdf estimation from an arbitrary number of images. ACM Trans. Graph. **38**(4), 134:1–134:15 (2019)
11. Gardner, A., Tchou, C., Hawkins, T., Debevec, P.: Linear light source reflectometry. ACM SIGGRAPH 2003 Papers, SIGGRAPH '03 pp. 749–758 (2003)
12. Hui, Z., Sunkavalli, K., Lee, J.Y., Hadap, S., Wang, J., Sankaranarayanan, A.C.: Reflectance Capture Using Univariate Sampling of BRDFs. Proceedings of the IEEE International Conference on Computer Vision **2017-Octob**, 5372–5380 (2017)
13. Ikehata, S.: Cnn-ps: Cnn-based photometric stereo for general non-convex surfaces. In: The European Conference on Computer Vision (ECCV) (2018)

14. Kang, K., Chen, Z., Wang, J., Zhou, K., Wu, H.: Efficient reflectance capture using an autoencoder. ACM Transactions on Graphics **37**(4) (2018)
15. Kang, K., Xie, C., Yi, M., Gu, M., Chen, Z., Zhou, K., He, C., Wu, H.: Learning Efficient Illumination Multiplexing for Joint Capture of Reflectance and Shape **38**(6) (2019)
16. Kazhdan, M., Hoppe, H.: Screened poisson surface reconstruction. ACM Trans. Graph. **32**(3) (2013)
17. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Y. Bengio, Y. LeCun (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015)
18. Li, X., Dong, Y., Peers, P., Tong, X.: Modeling surface appearance from a single photograph using self-augmented convolutional neural networks. ACM Trans. Graph. **36**(4), 45:1–45:11 (2017)
19. Li, Z., Sunkavalli, K., Chandraker, M.: Materials for masses: Svbrdf acquisition with a single mobile phone image. In: V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss (eds.) Computer Vision – ECCV 2018, pp. 74–90. Springer International Publishing, Cham (2018)
20. Li, Z., Xu, Z., Ramamoorthi, R., Sunkavalli, K., Chandraker, M.: Learning to reconstruct shape and spatially-varying reflectance from a single image. SIGGRAPH Asia 2018 Technical Papers, SIGGRAPH Asia 2018 **37**(6) (2018)
21. Lu, F., Chen, X., Sato, I., Sato, Y.: Symps: Brdf symmetry guided photometric stereo for shape and light source estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(1), 221–234 (2018)
22. Lu, F., Matsushita, Y., Sato, I., Okabe, T., Sato, Y.: Uncalibrated photometric stereo for unknown isotropic reflectances. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2013)
23. Matusik, W., Pfister, H., Brand, M., McMillan, L.: Efficient isotropic brdf measurement. In: Proceedings of the 14th Eurographics Workshop on Rendering, EGRW '03. Eurographics Association (2003)
24. Nam, G., Lee, J.H., Gutierrez, D., Kim, M.H.: Practical SVBRDF acquisition of 3D objects with unstructured flash photography. SIGGRAPH Asia 2018 Technical Papers, SIGGRAPH Asia 2018 **37**(6) (2018)
25. Nehab, D., Rusinkiewicz, S., Davis, J., Ramamoorthi, R.: Efficiently combining positions and normals for precise 3d geometry. In: ACM SIGGRAPH 2005 Papers, SIGGRAPH '05, p. 536–543 (2005)
26. Nielsen, J.B., Jensen, H.W., Ramamoorthi, R.: On optimal, minimal BRDF sampling for reflectance acquisition. ACM Transactions on Graphics **34**(6) (2015)
27. Oxholm, G., Nishino, K.: Shape and Reflectance Estimation in the Wild. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(2), 376–389 (2016)
28. Papadhimitri, T., Favaro, P.: A closed-form, consistent and robust solution to uncalibrated photometric stereo via local diffuse reflectance maxima. Int. J. Comput. Vision **107**(2), 139–154 (2014)
29. Ren, P., Wang, J., Snyder, J., Tong, X., Guo, B.: Pocket reflectometry. ACM Transactions on Graphics **30**(4), 1–10 (2011)
30. Riviere, J., Peers, P., Ghosh, A.: Mobile Surface Reflectometry. Computer Graphics Forum **35**(1), 191–202 (2016)
31. Santo, H., Samejima, M., Sugano, Y., Shi, B., Matsushita, Y.: Deep photometric stereo network. In: 2017 IEEE International Conference on Computer Vision Workshop (ICCVW). IEEE Computer Society (2017)
32. Shi, B., Matsushita, Y., Wei, Y., Xu, C., Tan, P.: Self-calibrating photometric stereo. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2010)
33. Shi, B., Mo, Z., Wu, Z., Duan, D., Yeung, S., Tan, P.: A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(2), 271–284 (2019)
34. Walter, B., Marschner, S., Li, H., Torrance, K.: Microfacet models for refraction through rough surfaces. Eurographics pp. 195–206 (2007)
35. Ward, G.J.: Measuring and modeling anisotropic reflection. In: Proceedings of the 19th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '92. Association for Computing Machinery (1992)
36. Wu, H., Zhou, K.: AppFusion: Interactive Appearance Acquisition Using a Kinect Sensor. Computer Graphics Forum **34**(6), 289–298 (2015). DOI 10.1111/cgf.12600
37. Wu, Z., Tan, P.: Calibrating photometric stereo by holistic reflectance symmetry analysis. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2013)
38. Xu, Z., Nielsen, J.B., Yu, J., Jensen, H.W., Ramamoorthi, R.: Minimal BRDF sampling for two-shot near-field reflectance acquisition. ACM Transactions on Graphics **35**(6) (2016)
39. Xu, Z., Sunkavalli, K., Hadap, S., Ramamoorthi, R.: Deep image-based relighting from optimal sparse samples. ACM Transactions on Graphics **37**(4) (2018)
40. Zhou, K., Synder, J., Guo, B., Shum, H.Y.: Isocharts: stretch-driven mesh parameterization using spectral analysis. In: Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing, pp. 45–54 (2004)