Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

On-Site Single Image SVBRDF Reconstruction with Active Planar Lighting

Anonymous for Review

ARTICLE INFO

Article history:
Received April 18, 2025

Keywords:
Appearance Modeling
SVBRDF
Planar Lighting
Uncontrolled environments

ABSTRACT

Recovering the spatially-varying bidirectional reflectance distribution function (SVBRDF) from a single image in uncontrolled environments is challenging while essential for various applications. In this paper, we address this highly ill-posed problem using a convenient capture setup and a carefully designed reconstruction framework. Our proposed setup, which incorporates an active extended light source and a mirror hemisphere, is easy to implement for even common users and requires no careful calibration. These devices can simultaneously capture uncontrolled lighting, real active lighting patterns, and material appearance in a single image. Based on all captured information, we solve the reconstruction problem by designing lighting clues that are semantically aligned with the input image to aid the network in understanding the captured lighting. We further embed lighting clue generation into the network's forward pass by introducing real-time rendering. This allows the network to render accurate lighting clues based on predicted normal variations while jointly learning to reconstruct high-quality SVBRDF. Moreover, we also use captured lighting patterns to model noises of pattern display in real scenes, which significantly increases the robustness of our methods on real data. With these innovations, our method demonstrates clear improvements over previous approaches on both synthetic and real-world data.

1. Introduction

Real-world material reconstruction is a long-standing problem in computer graphics, with wide-ranging applications like photorealistic rendering. Traditional methods [1, 2] that rely on custom devices to capture material in the 4D domain are time-consuming and restricted to laboratory settings.

Recently, lightweight material capture and reconstruction from very sparse images using deep learning have attracted significant attention. This problem is highly ill-posed because of insufficient measurements. To reduce the ambiguities between material and lighting, many methods restrict the capture setup to contain only active light sources, such as point lighting [3, 4, 5], linear lighting [6], or planar lighting [7]. These setups allow networks to be aware of the simple highlight response of the fixed light source, thereby improving reconstruction accuracy. Yet, the strict assumption of a dark room severely limits the applicability of these methods.

Contrastingly, capturing material in natural scenes is more desirable for common users. Uncontrolled lighting, however, can cause unstable appearances that limit reconstruction performance. Through capturing the lighting, Lin et. al.[8] can reconstruct the material of non-planar exemplars while assuming homogeneous BRDF. Instead, most methods assume unknown lighting to avoid the additional burden of capturing uncontrolled lighting. Several methods [9, 10] try to reconstruct under only environment lighting, while limiting the specular component of captured material to reduce ambiguities. A more promising way is adding active lighting in uncontrolled environments to activate relatively stable appearances. Many methods [11, 12, 13, 14, 15] choose the co-located flash lighting because of its flexibility. However, it is still challenging to reconstruct accurate SVBRDFs from one or two shots under unknown lighting.

In this paper, we propose a convenient capture setup and a

light-aware framework to reconstruct SVBRDFs from a single image captured in uncontrolled environments. Our setup needs no careful calibration and enables hand-held capture of both material appearance and incident lighting within a single shot, while our framework comprehensively interprets all captured information to reconstruct high-quality SVBRDFs.

Specifically, we extend the planar lighting setup [7] originally designed for darkroom conditions to uncontrolled environments by introducing a tiny mirror ball. In this new setting, the planar light source plays a crucial role in occluding ambient light along the specular reflection direction, ensuring stable excitation of material specularities on planar SVBRDFs. Meanwhile, the mirror ball enables joint capture of environment illumination, active light patterns, and material appearance, allowing us to model uncontrolled illumination and display noise during real-world data capture and effectively address key ambiguities in reconstruction.

To leverage the captured environment lighting in reconstruction, we design a type of lighting clue that is semantically aligned with the captured appearance. Technically, the lighting clue is appearance rendered with pre-defined basic materials under the captured environment lighting. This alignment bridges the appearance and illumination domains, which effectively reduces the learning burden on the reconstruction network. To ensure the lighting clue reflects the uncontrolled illumination in a normal-aware way, we dynamically render it during training using the predicted surface normal, enabling tight coupling with the appearance and enhancing reconstruction quality. Besides, based on the captured active lighting pattern, we introduce a denoising module to address the discrepancy between the ideal lighting pattern in training and the noisy lighting pattern in real-world settings. It models display- and capture-induced noise of lighting patterns in real data, and then effectively maps captured appearances to their ideal synthetic counterparts, improving the robustness of reconstruction on real-world data.

We demonstrate our method on both synthetic and real data including various materials. In summary, our method shows superior performance than previous methods thanks to the following contributions:

- a simple setup that not only can efficiently activate material specular responses, but can capture uncontrolled lighting and displayed active patterns jointly in a single image;
- a light-aware reconstruction framework that jointly leverages captured environmental lighting, active patterns, and material appearance for high-quality SVBRDF recovery;
- a semantically appearance-aligned, dynamically rendered lighting clue to bridge appearance and illumination during training.

2. RELATED WORK

In this section, we focus on recent methods that capture material from sparse images. These methods are classified into two categories: those that only rely on active lighting and those that capture materials under uncontrolled lighting. For a more comprehensive discussion about previous methods, we refer readers to several excellent surveys [16, 17, 18].

2.1. Capture Material with Only Active Lighting

Material reconstruction from sparse images is a highly ill-posed problem. To address this challenge, a majority of methods restrict lighting to only active lighting setups to reduce ambiguities in appearance. Since point lighting can provide various incident lighting directions across surfaces, many methods attempt to reconstruct material from just a single flash image. For instance, based on the assumption of statistical materials, Aittala et al. [12] proposed using a neural texture descriptor to optimize the similarity between target material maps and different patches within the input image. Leveraging this strong material assumption, several recent methods [19, 20, 21] are capable of generating materials in an unsupervised manner. However, to handle more general material properties, Deschaintre et al. [3] and Li et al. [22] built synthetic datasets respectively to train networks for SVBRDF map estimation. Following the success of deep learning, many methods are proposed to explore various priors through learning, such as soft highlight mask supervision [4], meta-learned priors for test-time optimization [23], real-data distribution modeling [24, 25], basis-material assumption [5], and frequency domain decomposition [26]. In addition, Guo et al. [27] reconstructed materials from a single high-resolution flash image by combining local prediction and global fusion using attention mechanisms. More recently, several methods have explored problem break-down [28], learned Gradient Descent [29] and correlation perception [30] to model SVBRDF reconstruction from a single flash image. Compared with point lighting, Zhang et al. [7] introduced planar lighting to improve capture efficiency for single-image material reconstruction and achieve impressive accuracy. As the most related work to ours, their setup inspires our capture design. However, we extend it with a key addition—a mirror ball—to simultaneously capture environment lighting and active light patterns. Moreover, their method cannot be directly adapted to uncontrolled environments, as it lacks the ability to model appearance variations caused by unknown illumination and is sensitive to active lighting display noise in real-world settings.

In the context of multiple inputs, Riviere et al. [31] proposed a method that can reconstruct materials from a mounted LCD screen with gradient illumination using polarization clues. Hwang et al. [32] also used polarization and designed a hand-held manner to capture Polarimetric SVBRDF [33]. With only flash images, Nam et al. [34] capture material and geometry of 3D objects yet needs hundreds of images. When only sparse images are available, many methods utilizing the learned data prior to reconstruct SVBRDF of planar objects. Gao et al. [35] introduced a method to optimize within a latent material space through an analysis-by-synthesis process. Building on this, Guo et al. [36] further leveraged a Generative Adversarial Network to serve as a more robust latent space. Although these methods perform well with sparse inputs, they all require complex registration or careful calibration. To solve this limitation, Deschaintre et al. [37] and Zhu et al. [38] proposed solutions that

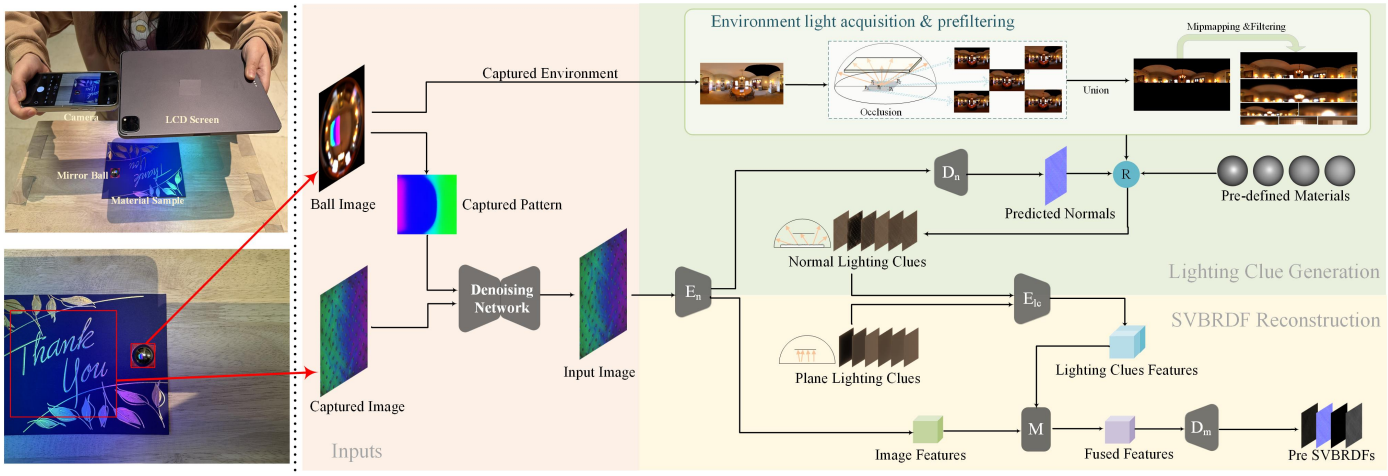


Fig. 1. Overview of the capture setup and two-branch network. The device which includes the material and a mirror ball, is placed in relative positions to capture a photo under composite lighting. On real data, the photograph is cropped and denoised to provide the network input, whereas denoise networks are pre-trained with perturbed patterns and appearance images. In the nLCs generation branch, the mirror ball is unfolded to obtain an environment map, which is processed through occlusion modeling, pre-convolution, and mipmapping. The appearance image is fed into the encoder and normal decoder to predict normals, which are combined with basis materials to generate nLCs via real-time rendering. In the SVBRDF reconstruction branch, features from nLCs, pLCs, and the appearance image are fused in the intermediate module and passed to the decoder for material reconstruction.

use two smartphones yet require a tripod to mount the camera. Recently, Wang et al. [39] design a near-far field capture method that achieves impressive results with two shots under only a flash. Different from these methods that are based on off-the-shell devices, Xian et al.[40] construct a stable illumination environment with the setup to reconstruct planar SVBRDFs from multiple images, but its setup required custom fabrication.

However, all these methods only utilize active lighting for material capture, which significantly limits their application in daily life. For common users, capturing materials on-site is far more desirable while uncontrolled lighting is challenging remains a major challenge to current methods.

2.2. *Capture Material in the Wild*

Another class of methods aims to capture materials in uncontrolled environments. Due to the complex entanglement between material and lighting, these methods either rely on additional inputs to obtain more appearance measurements or impose strong assumptions on the captured materials. Aittala et al. [11] pioneered this problem by capturing flash and no-flash image pairs while assuming that the material sample is statistically stationary. Through data augmentation, Li et al. [9] and Ye et al. [41] trained networks to reconstruct materials from a single image captured in the wild but only estimate homogeneous specular terms. By separately capturing the light, Lin et al. can reconstruct high-quality material with mesostructure from a single HDR photo, yet assuming the material is homogeneous. Rivire et al. [42] can reconstruct full SVBRDF maps with only three rotation of linear polarizing filter, while they need complex calibration for high-quality reconstruction. Besides, Martin et al. [10] and Li et al. [43] also assumed materials are non-metallic to simplify the problem. This simplification allows them to address the problem of reconstructing high-resolution material maps and decomposing non-planar objects. Given a flash/no-flash pair and through deep learning, Boss et al. [13] can further estimate specular maps. However, the unstable appearance caused by unknown lighting is still challenging to their method. Instead, our method can jointly capture environment lighting and appearance in a single image, which greatly reduces the ambiguities.

Recently, several methods [14, 44] utilized the general image data priors by fine-tuning the Stable Diffusion model [45] to learn a specific material prior. By using captured images under uncontrolled lighting as conditions for generation, these methods can sample the diffusion model to reconstruct various plausible material maps for users. However, their results are designed to prioritize diversity rather than achieving accurate reconstruction of the input appearance images.

In comparison, our method has no assumption on captured material and only requires a single image. Based on the convenient setup and carefully designed lighting clues, our method can reconstruct high-quality SVBRDF maps of various materials under controlled environments.

3. METHOD

3.1. *Problem Formulation And Method Overview*

We aim to estimate high-quality SVBRDFs from a single image captured on-site under environment lighting. We assume that the reflectance properties at each surface point of the planar material can be well represented by the Cook-Torrance [46] BRDF model with a GGX [47] surface distribution. Based on this assumption, the SVBRDF at each point x on the surface can be expressed as $s(x) := (k_s, \rho, k_d, n)$ defined by four material parameters: specular albedo k_s, roughness ρ, diffuse albedo k_d and surface normal n. However, the pixel values of the captured appearance image are inherently ambiguous because of the non-linearity between lighting and material properties. This ambiguity becomes even more pronounced when images are

captured under uncontrolled lighting conditions, making the reconstruction problem significantly more challenging.

To reduce ambiguities, we introduce a planar light source to stably capture reflectance responses, which can be easily implemented in daily life. In this configuration, each captured pixel value $I(x, \omega_o)$ represents the radiance of a surface point x on the material sample, which can be represented as the following formula:

$$I(x) = \int_{C_\Omega^P} L_i^e(\omega_i) D(\omega_i, s) d\omega_i + \int_P L_i^p(x, \omega_i) D(\omega_i, s) d\omega_i, \quad (1)$$

where $D = f_r(\omega_i, \omega_o, s)(n \cdot \omega_i)$, $f_r(\cdot)$ is the BRDF function, ω_i is the incident directions and ω_o is the view directions. The near-field planar light source is modeled as a polygon P with the lighting pattern $L_i^p(\cdot)$ while the far-field environment lighting is described as a sphere map $L_i^e(\cdot)$ over the upper hemisphere Ω. Additionally, C_Ω^P represents the remaining hemispherical area caused by the occlusion from the planar light source.

To reconstruct materials from such images, we propose a hand-held setup that can capture appearance together with all incident lighting (Sec. 3.2) and a deep learning-based framework that can reconstruct high-quality SVBRDFs by comprehensively understanding lighting conditions, as shown in Fig. 1. The framework consists of an reconstruction network and a denoise network. The reconstruction network reasons about the captured passive incident lighting to facilitate accurate material estimation, while the denoise network filter the capture noise in appearance images based on the captured active incident lighting.

Given the passive environment lighting, we propose a form of lighting clues (Sec. 3.3) to aid reconstruction network in lighting understanding. The lighting clues are generated by rendering a predefined set of basis materials $\{s_i^b := (k_{si}^b, r_i^b, k_{di}^d, n_i^b) \mid i = 1, ..., k\}$ under the captured lighting conditions. Similar to the input image, each pixel in the lighting clues corresponds to a reflectance response of the same surface point as the input image yet rendered with known material properties. This semantic alignment is beneficial for SVBRDF reconstruction using a convolution neural network. Nevertheless, as shown in Fig. 2, normal is critical for lighting clues, since directly using plane normal $\overline{n}_i^b$ for lighting clue rendering (denoted as $\overline{I}_i^{lc}$) only capture lighting concentrated at top areas. This limitation reduces the ability to model accurate appearance at the edge. We address this by using predicted normals $\hat{n}_i^b$ in lighting clue rendering (denoted as $\hat{I}_i^{lc}$) to render more accurate appearance information. Thus, we embed the lighting clue rendering into the forward pass of reconstruction network G. This allows the network to automatically learn globally optimal lighting clues, further improving reconstruction accuracy.

As a result, our reconstruction network G (Sec. 3.4) consists of two branches: the lighting clue generation branch and the SVBRDF reconstruction branch. In the former branch, we first predict an initial normal and introduce the split-sum technique [48] to render $\hat{I}_i^{lc}$ on the fly, which are then encoded into features and subsequently fused in the latter branch for SVBRDF reconstruction. To deal with the occlusion caused by near-field planar lighting, which is challenging for split-sum, we propose

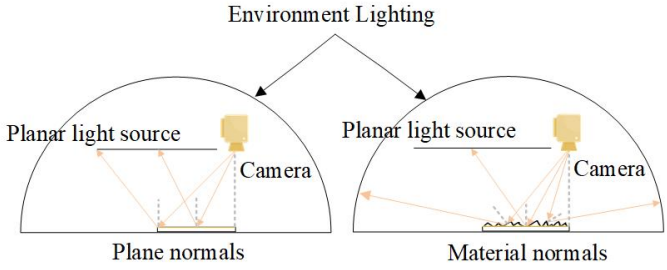


Fig. 2. Illustration of plane normals and material normals. The hemisphere represents the environment lighting around the material, containing the planar light source and the camera. The dashed line indicates the normal at a specific point on the material, while the arrow shows the direction of the reflected light.

to mask the sphere map using representative surface points. The $\overline{I}_i^{lc}$ are used to replenish the missing information in the masked lighting. Hence, the training process can be described as optimizing the network parameters θ and minimizing the loss function $\mathcal{L}(\cdot, \cdot)$ as follows:

$$\theta_* = \arg\min_\theta \mathcal{L}(G(I,\ \overline{I}_i^{lc}(\overline{s}_i^b),\ \hat{I}_i^{lc}(\hat{s}_i^b)),\ s_{gt}; \theta), \quad (2)$$

where s_{gt} is the ground-truth material maps, $\hat{s}_i^b$ consists of predefined materials and the predicted normal $\hat{n}_i^b$ generated during the forward pass of $G(\cdot)$, while $\overline{s}_i^b$ uses plane normal $\overline{n}_i^b$.

Besides, the reconstruction network is trained to disentangle material properties based on an ideal active lighting pattern. However, in real-world captures, the active lighting pattern may deviate from its ideal counterpart due to noise introduced by display and capture devices. To address this discrepancy, we leverage the captured active lighting pattern and design a denoising network (Sec. 3.4) that explicitly models the dominant sources of display-induced noise. By learning this noise model, the denoising network transforms the noisy captured appearance into a clean version corresponding to the ideal active lighting pattern, thereby improving the robustness of our reconstruction on real-world data.

After the training of all networks, the testing process begins with capturing a single image that includes both the mirror ball and the material. From this image, we detect the position of the mirror ball and unfold it to obtain the environment lighting map and active lighting patterns. The environment map is then pre-processed to generate plane lighting clues for network input and lighting mipmaps for split-sum rendering. Meanwhile, the material appearance is automatically cropped from the captured image based on a setup-derived Field-of-View (FOV) setting. Together with the active lighting patterns, the cropped appearance is passed through the denoise network to remove display noise. Subsequently, the denoised appearance and lighting clues are input into the SVBRDF reconstruction network for material estimation.

3.2. Acquisition Setup

The acquisition setup we proposed in the method is based on the device introduced by Zhang et al [7]. They defined the relative positions of the LCD screen, camera, and material sample.

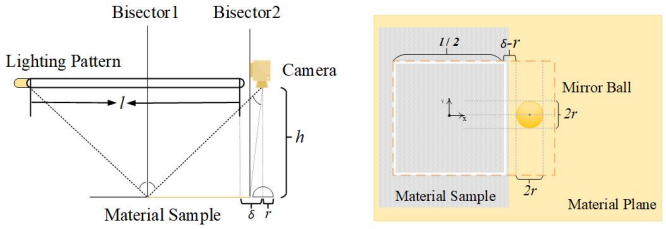


Fig. 3. Side and top views of the capture setup. We added a mirror ball along the camera's optical axis to the existing geometric setup and designed a material sample area and a mirror ball positioning method to simplify the capture process.

This relative geometry configuration requires no precise calibration and can be quickly implemented by common users. As shown on the left of Fig. 3, the user only needs to roughly place the screen and camera are approximately parallel to the material surface, while the camera is next to the middle of a screen edge. Assuming the screen has a size of $l \times l$, the appearance of a material sample of size $l/2 \times l/2$ can be automatically cropped through a specific FOV region derived from the height h. This allows our network to be trained with a relative geometry setting on synthetic data. Besides, a small distance δ between the camera and screen are introduced to consider the physical width of camera lens. Under this configuration, the optical axis of the camera is nearly aligned with the normal of the screen. Within the derived FOV range, each surface point is captured near the specular reflection direction.

However, only taking the appearance image as input is severely ambiguous to SVBRDF reconstruction under uncontrolled environments. Besides, over half FOV range of the captured image is wasted for capture in mirror directions. To handle on-site capturing, we introduce a tiny mirror hemisphere to the capture setup (Fig. 3). With this mirror ball, we can jointly capture the uncontrolled lighting and real active pattern while needing nearly no extra capture burden for users. The former lighting is a strong clue for networks to disentangle the environment contributions in appearance, and the latter lighting can model the pattern display perturbations for denoise which we will discuss later.

Specifically, we assume that the mirror ball is placed at the intersection of the camera's optical axis and the material plane, which is an easy positional reference for users to set up the device. Hence, the ball should be captured at the center of the input image on synthetic data. In the top view of the material plane (right side of Fig. 3), the distance between the mirror ball and the material sample can be calculated as $\delta - r$. By assuming the coordinate system origin is the center of the material sample area, we can easily calculate the circumscribed square of the mirror ball as follows:

$$Square := \left\{ \begin{bmatrix} \frac{l}{4} + \delta - r, r \end{bmatrix}, \quad \begin{bmatrix} \frac{l}{4} + \delta + r, r \end{bmatrix}, \atop \begin{bmatrix} \frac{l}{4} + \delta - r, -r \end{bmatrix}, \quad \begin{bmatrix} \frac{l}{4} + \delta + r, -r \end{bmatrix} \right\} \tag{3}$$

In real data capture, this ideal geometry relation from mirror ball to material sample will be used to separate the active lighting pattern and environment lighting. However, since the strict position of the mirror ball can not be promised, we use edge detection techniques on real data to accurately locate the position of the mirror ball. Given the detected FOV range of the mirror ball and the assumed geometry relation among device components, the active lighting pattern can be easily marked in the reflected lighting of the mirror ball. Users can decide whether to make subtle position adjustments based on the accuracy of pattern detection. We argue that it is not necessary to place the mirror ball exactly at the center of the input image. Since the captured environmental lighting is assumed to be at infinity, slight deviations in the mirror ball's position have negligible impact on passive lighting capture. For active lighting, the segmentation is initially based on the ideal geometric configuration, and any misalignment caused by position deviation can be easily corrected by the user through slight adjustments to the planar light's position and orientation. This refinement process is lightweight and typically takes less than one minute to complete. As a result, minor mirror ball positioning errors have minimal impact on reconstruction accuracy.

3.3. Lighting Clues Rendering

As mentioned before, we designed lighting clues that are semantically aligned to the sampled appearance images. For basis materials, we choose the most representative basis materials s_i^b by changing the values of roughness and adjusting the ratio of k_s and k_d while keeping their sums constant. Since there are significant variations in surface normal for most real materials, the $\bar{I}_i^{lc}$ rendered using upward-facing normal are not desirable to indicate environment lighting. Thus, we choose to use initial normal in basis materials for the render of $\hat{I}_i^{lc}$, which can match the appearance in captured images better.

One straightforward way to obtain $\hat{I}_i^{lc}$ is to use existing methods for initial normal prediction and then rendering with basis materials. However, this may result in error propagation of normal prediction into the network, which can negatively affect subsequent reconstruction. Therefore, we employed real-time rendering to make the network jointly learn initial normal prediction and SVBRDF estimation for global optimal reconstruction.

Real-time Rendering of $\hat{I}_i^{lc}$. We use the split-sum technique [48] to render $\hat{I}_i^{lc}$ online in training because of its rendering efficiency and frequency-free lighting processing, which can be represented below:

$$\hat{I}_i^{lc} = \underbrace{\int_\Omega \hat{L}_e^i(\omega_i) d\omega_i}_{Prefiltering} \underbrace{\int_\Omega D(\omega_i, \hat{s}_i^b) d\omega_i}_{LUT Fetching}, \tag{4}$$

where $\hat{L}_e^i$ is a sphere map with the occlusion of the active light source in our method. Specifically, we prefilter the sphere map and apply online Look-Up-Table (LUT) fetching using the predicted normal and basis materials.

Environment lighting Occlusion. It is worth noting that real-time rendering cannot accurately simulate the impact of environment lighting occlusion caused by the near-field active planar light source. To approximate the light received by material surface points, we modeled the near-field occlusion by pre-masking the sphere map using several surface points. As shown

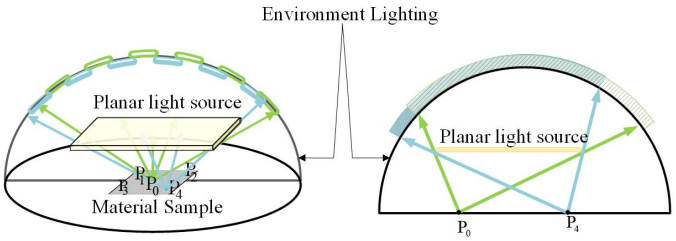


Fig. 4. Schematic diagram for modeling active lighting occlusion. The hemisphere represents the environment light, the rectangle represents the active light occlusion area, the green line indicates the light ray reaching surface point p_0 of the material sample, and the blue line indicates the light ray reaching point p_4. The received environment illumination of surface points is spatially varying.

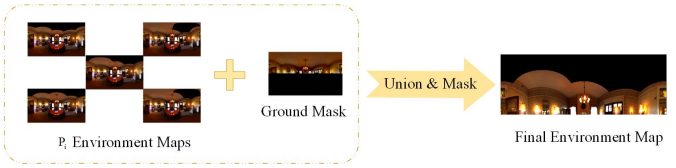


Fig. 5. In the left, we show the sphere maps of $\{p_i\}$ and hemisphere maps that are occluded individually, while the right is the unified sphere map.

in Fig. 4, the hemisphere, which is the entire environment illumination, is blocked differently by the active planar light source for different surface points.

To calculate the masked sphere map for prefiltering, we selected the five most representative surface points $p_i := \{p_0, p_1, p_2, p_3, p_4\}$, consisting of four corners and the center point, to approximate the entire occlusion range of the lighting rectangle. Based on the spatial relationship between the lighting rectangle and these points, we can model the occlusion as a set of directions:

$$\{\omega_i := \frac{x_l - x}{\|x_l - x\|} \mid x_l \in P \wedge x \in \{p_i\}\} \tag{5}$$

where, x_l represents a point on the active planar light source P, and x is a predefined P_i on the material sample. The term $x_l - x$ denotes the directions of light that intersect with the active planar light source region, corresponding to the occluded parts. Additionally, we also filter out the lighting of the lower hemisphere. Finally, as illustrated in Fig. 5, we can take the union of these environment maps and lower hemisphere mask to obtain the final environment map $\hat{L}_e^i$. This sphere map is used for prefiltering and the real-time rendering of $\hat{I}_i^{lc}$.

However, the mask area is excessive for ground-truth lighting transport of every surface point. As illustrated in the side view of Fig. 5, the approximated occlusion range is different and larger than that of both P_0 and P_4. The appearance clues of each surface point rendered with environment information outside the accurate occlusion range should be compensated to the network. Thus, we also use the $\overline{I}_i^{lc}$ as the inputs, which are correctly pre-rendered using path tracing to model near-field occlusion. In summary, the complete environment lighting information with near-field occlusion is stored in the combination of $\hat{I}_i^{lc}$ and $\overline{I}_i^{lc}$, as shown in the yellow box of Fig. 6.

Environment map process. Based on the masked sphere map

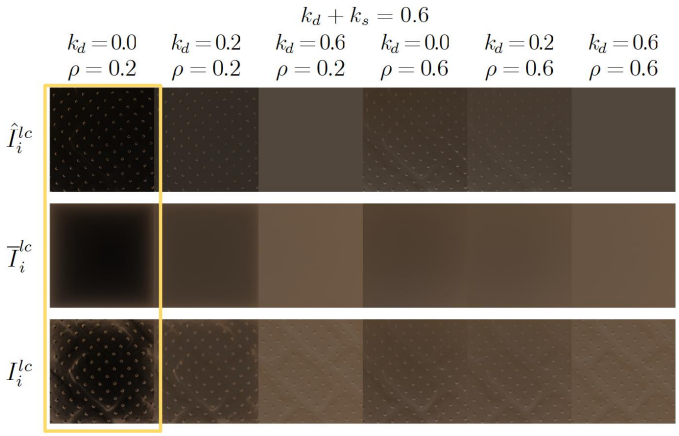


Fig. 6. The example results of all lighting clues and ground-truth lighting clues rendered with path-tracing. $\hat{I}_i^{lc}$ are rendered with predicted normal and occluded environment lighting, while $\overline{I}_i^{lc}$ are rendered with upward normal. In the third row, I_i^{lc} are rendered with ground-truth normal and path tracing to model spatially-varying occlusion of the near-field active light source.

with occlusion modeling, we apply the pre-convolution and mipmap techniques to process the $\hat{L}_e^i$ according to the split-sum. As shown in Fig.7, we prefilter the sphere map using Monte-Carlo sampling and generate several LoDs with mipmapping [49]. This process approach can reduce the computational cost of the lighting integral, and the rendering of environment lighting becomes two simple texture-fetching processes based on materials.

3.4. Network Structure and Loss Function

In this section, we will introduce in detail our two-branch network G and denoise network $G_{denoise}$ along with the associated supervisions, and first provide a brief overview of their designs.

During training on synthetic data, the inputs of G are an appearance image I rendered under both passive and active lighting, the prefiltered mipmap stacks of environment map $\hat{L}_e^i$, and plane lighting clues $\overline{I}_i^{lc}$. Based on these inputs, G predicts material properties by first rendering normal lighting clues and then fusing all information to produce the final SVBRDF maps. To account for pattern display noise present in real-world data, we simulate the primary noise sources during training and generate a noisy version of the appearance I'_0, corresponding to the real captured input. The denoise network $G_{denoise}$ is designed to filter I'_0 and obtain the final input image I, which is then fed into G. We next detail the architectures and designs of these two networks.

Two-Branch Network. Given the input image I, a shared feature vector is first extracted by the encoder E_n, resulting in $F_I = E_n(I)$. Based on this shared representation, the two branches are subsequently processed in parallel.

In light clue generation branch, we use prefiltered environment maps and split-sum to render $\hat{I}_i^{lc}$ through initial normal map prediction. Specifically, the F_I is fed into the normal decoder $D_n(\cdot)$ to predict initial normals:

$$\hat{n} = D_n(F_I). \tag{6}$$

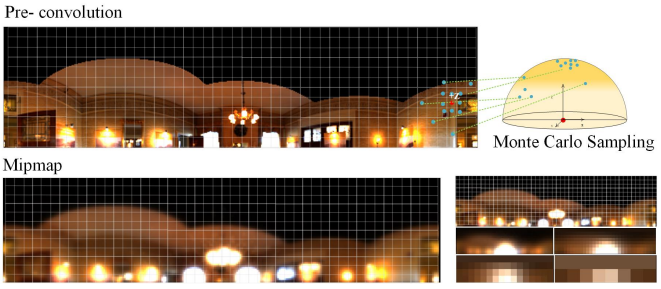


Fig. 7. Environment map processing. The sphere map with occlusion modeling is processed with Monte Carlo sampling. The maps are prefiltered according to split-sum and mipmaps are generated to store prefiltered lighting at different roughness scales, which are used by simple texture queries in real-time rendering of $\hat{I}_i^{lc}$.

Following, the predicted normals are combined with predefined basis materials s_b to serve as the input of real-time renderer. Given the prefiltered mipmaps of lighting with occlusion, $\hat{I}_i^{lc}$ can be rendered using Eq. 4. After real-time rendering, $\hat{I}_i^{lc}$ provide a foundation for material reconstruction, which hints appearance rendered under the captured environment lighting to the network. However, we observed that split-sum and our occlusion modeling will cause color bias in $\hat{I}_i^{lc}$ compared with that rendered with path tracing. Thus, we convert $\hat{I}_i^{lc}$ into gray-scale before the feature encoding and fusion. This conversion removes color bias that could potentially interfere with the reconstruction process while retaining the essential geometric and lighting information. Furthermore, the gray-scale representation can force the network to focus on the structural information which is critical for accurate reconstruction.

In the SVBRDF reconstruction branch, all lighting clues are encoded into latent space and are fused with the feature of the input image to reconstruct accurate SVBRDFs. Specifically, we simply concatenate the gray-scale $\hat{I}_i^{lc}$ generated by the former branch and the pre-rendered $\bar{I}_i^{lc}$ and encode them with an encoder $E_{lc}(\cdot)$ to generate features F_{lc}. Then combine F_{lc} with F_I and pass them to the intermediate block $M(\cdot)$ to generate fused features F_{all}.

$$F_{lc} = E_{lc}(\hat{I}_i^{lc}, \bar{I}_i^{lc}), \qquad F_{all} = M(F_I, F_{lc}) \tag{7}$$

The $M(\cdot)$ is a stack of base network blocks to progressively extract and fuse the input features. Finally, the F_{all} are fed into the material reconstruction decoder $D_s(\cdot)$ to estimate SVBRDFs $s = D_m(F_{all})$. Both network branches are trained jointly to achieve global optimal lighting clue generation and SVBRDF reconstruction.

Denoise network. Due to the variability of display devices and synthetic configuration, the patterns of the active light source L_i^a (Ideal Pattern) for input images rendering on synthetic data can have various noises under real-world conditions. Such non-linearity noise is difficult to learn directly. Fortunately, our introduced mirror ball can capture the real displayed lighting pattern in the single image. So we can model the noises based on the difference between real pattern and ideal pattern in synthetic rendering. This base color pattern can significantly reduce the learning burden of the denoise network. Note that, acquiring the real pattern has no extra capture burden for users since it is jointly captured by our mirror ball in the captured photograph. Overall, we address the learning of noise filtering by jointly inputting the noised pattern and noised image to the denoise network.

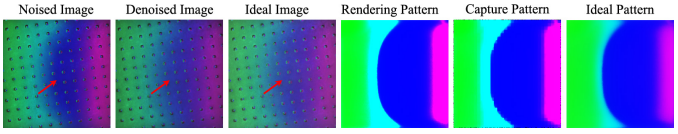


Fig. 8. An example of the input noised image and output denoised image of our denoise module. Noised images are rendered by the Rendering Pattern which is injected with three types of color noises from Ideal Pattern. Ideal Image are rendered by Ideal Pattern. The Capture Pattern has two more types of noises than the Rendering Pattern.

To achieve this, we design three types of noise on the ideal pattern L_i^a to simulate the real capture process: lighting intensity, non-linear color display, and white balance. These noises can closely simulate the lighting display perturbations in various real capture setups. Then, the noised input image I'_0 is rendered under the noised lighting pattern (Rendering Pattern) and GT material renderings. Similar to the noise in input image rendering, the resolution of the captured image and material properties of the mirror ball can also interfere with the captured active light pattern. To simulate this, we inject two additional types of noise into the lighting pattern: structural interference (mosaic effect) and random boundary erosion. Then, we input the final noised lighting pattern $\tilde{L}_i^a$ (Captured Pattern) to the denoise network. We show an example of noised images and patterns in Fig. 8. The input image I under the ideal synthetic rendering configuration can be obtained through the forward pass of the denoise network:

$$I = G_{denoise}(I'_0, \tilde{L}_i^a), \tag{8}$$

where I'_0 is a noised input image in training or an auto-cropped photograph in real-data capturing. By incorporating the denoise module, we enable users to conduct experiments using everyday display devices and ordinary mirror balls, significantly enhancing the practicality and accessibility of the setup.

Loss Functions. To supervise our two-branch network, our loss function consists of two parts:

$$\mathcal{L}_{train} = \lambda_1 L_n(\hat{n}, n_{gt}) + \lambda_2 L_r(s, s_{gt}). \tag{9}$$

The L_n term uses L1-norm loss to supervise the error between the predicted initial normal $\hat{n}$ and the ground truth normal n_{gt} in the lighting clue generation branch. The L_r term means using L1-norm loss to supervise the quality of the reconstructed SVBRDFs in the second branch. The weights $\lambda_1 = 1.0$ and $\lambda_2 = 1.0$ indicate the respective contributions of the two loss terms.

4. Implementation

In this section, we report our dataset and discuss some details of network design together with training hyper-parameters.

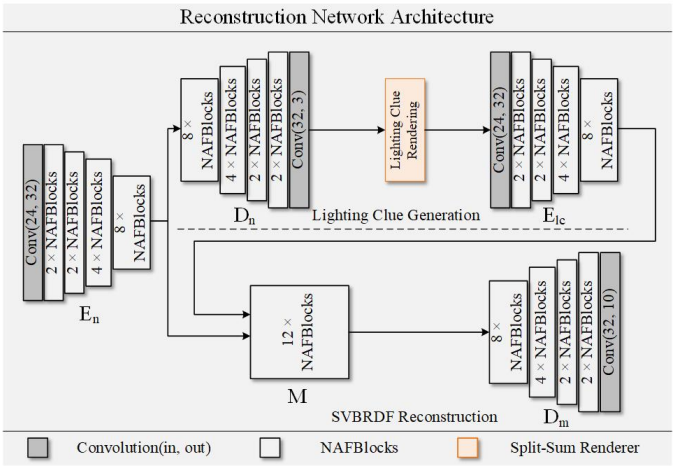


Fig. 9. Detailed architecture of our reconstruction network.

4.1. Network Details.

The base blocks of our network are based on NAFBlocks [50], which can address image-to-image translation problems efficiently. We then design the encoder and decoders to implement our reconstruction method. The resolution of our input and output images is 256×256. As shown in Fig. 9, the network begins with an initial convolution layer of 32 channels, followed by a 4-layer encoder with [2, 2, 4, 8] blocks per-layer, generating image features. In the lighting clue generation branch, image features are then passed through a 4-layer decoder with [2, 2, 2, 2] blocks per layer to predict the normal. For basis materials, we selected roughness values ($\rho = 0.2, \rho = 0.6$) and ($k_d := \{x \mid x \in \{0.0, 0.2, 0.6\}, k_s = 0.6 - k_d\}$) while the environment map is processed by prefiltering and mipmapping to generate 8 levels of mipmaps. Thus, there are 6 appearance images in $\hat{I}_i^{lc}$. In the SVBRDF reconstruction branch, 12 concatenated lighting clues ($\hat{I}_i^{lc}$ and $\overline{I}_i^{lc}$) are passed into a 4-layer encoder to extract clue features, which are combined with appearance features in a fusion module containing 12 blocks. The fused features are passed through a 4-layer decoder (2 blocks per layer) to predict final SVBRDF maps.

4.2. Training Details.

Lighting Pattern. In our method, active lighting pattern are essential for SVBRDF reconstruction. We fix the lighting pattern in training and choose the optimized one of Zhang et. al. [7], which aligns with our requirements and shows desirable accuracy on normal reconstruction.

Training Dataset. Our training dataset is based on MatSynth dataset [15], which is an artist-made and high-quality SVBRDF dataset for planar objects. Given the geometric configuration of the planar light source, planar object and camera in our setup, we simulate real captured process by rendering a planar object with different SVBRDFs from the training dataset under both planar lighting and environment lighting using a physically-based render engine—— Mitsuba3 [51]. For the environment lighting, we collected 256 sphere maps and randomly choose for each entity rendering. To alleviate the rendering burden of dataset preparation, we randomly selected 290, 700 material images and pre-render all input images for training. During the process of rendering data, we utilized an NVIDIA GeForce RTX 3090 card and spent 40 hours rendering the training dataset.

Hyper-parameters. We implemented our framework based on Pytorch and trained all networks with Adam Optimizer [52]. For two-branch network, the initial learning rate of the optimizer (β_1 =0.9, β_2 =0.9) is $1e-3$, which is gradually decrease to $1e-7$ using the cosine annealing algorithm. We trained both the two-branch network and denoise network for $400k$ iterations with a batch size of 6, respectively. Then, two-branch network are finetuned with add-on denoise network for $20k$ steps with the learning rate of $1e-5$. The training process spent 51 hours on an NVIDIA GeForce RTX 3090 card in total.

5. Experiments

In this section, we will compare our results with state-of-the-art methods on synthetic and real data to evaluate our method. Specifically, we compare our method with MatFusion[14], DeepBasis[5], and LPL[7]. MatFusion which has three variants is currently the best method for reconstructing SVBRDFs under environment lighting. For comparison, we selected its best-performing variant, the Two-shot method. DeepBasis and LPL achieve state-of-the-art reconstruction results from a single image captured in a darkroom under point light and the planar light source respectively. For a fair comparison, we retrained these methods using images captured under environment lighting and denoted the retrained versions as Deepbasis+ and LPL+, respectively. Similar to DeepBasis, recent methods [29, 28, 30] based on a single flash image also suffer from severely unstable reflectance responses caused by uncontrolled environment lighting. Besides, optimization [29] under unknown lighting is highly ill-posed while the training code is not provided by regression methods [28, 30]. Hence, we choose to compare with DeepBasis for fairness.

In ablation studies, we firstly verify our method by analyzing the impact of lighting clues form and different lighting clues' components on SVBRDF reconstruction results. Besides, regarding the setup, we analyze the robustness of our method by simulating different capture perturbations on synthetic data. Finally, we quantitatively demonstrate the generalization of the denoise model using a real displayed pattern. We further qualitatively compare the reconstruction results of our method with and without denoising on real-world data, and additionally perform comparisons with LPL+ using our denoised input images.

5.1. Comparative Experiments.

Results on Synthetic Data. We use the Root Mean Square Error (RMSE) to quantify the numerical differences between the predicted results and ground truth across all methods. Besides, we also use the perceptual similarity metric LPIPS [53] to evaluate re-rendered images from the perspectives of global structure, texture consistency, and visual similarity. The quantitative results are reported in Tab.1. The reconstructed reflectance properties of our method are superior to other methods, especially in terms of normals, roughness, and specular highlights.

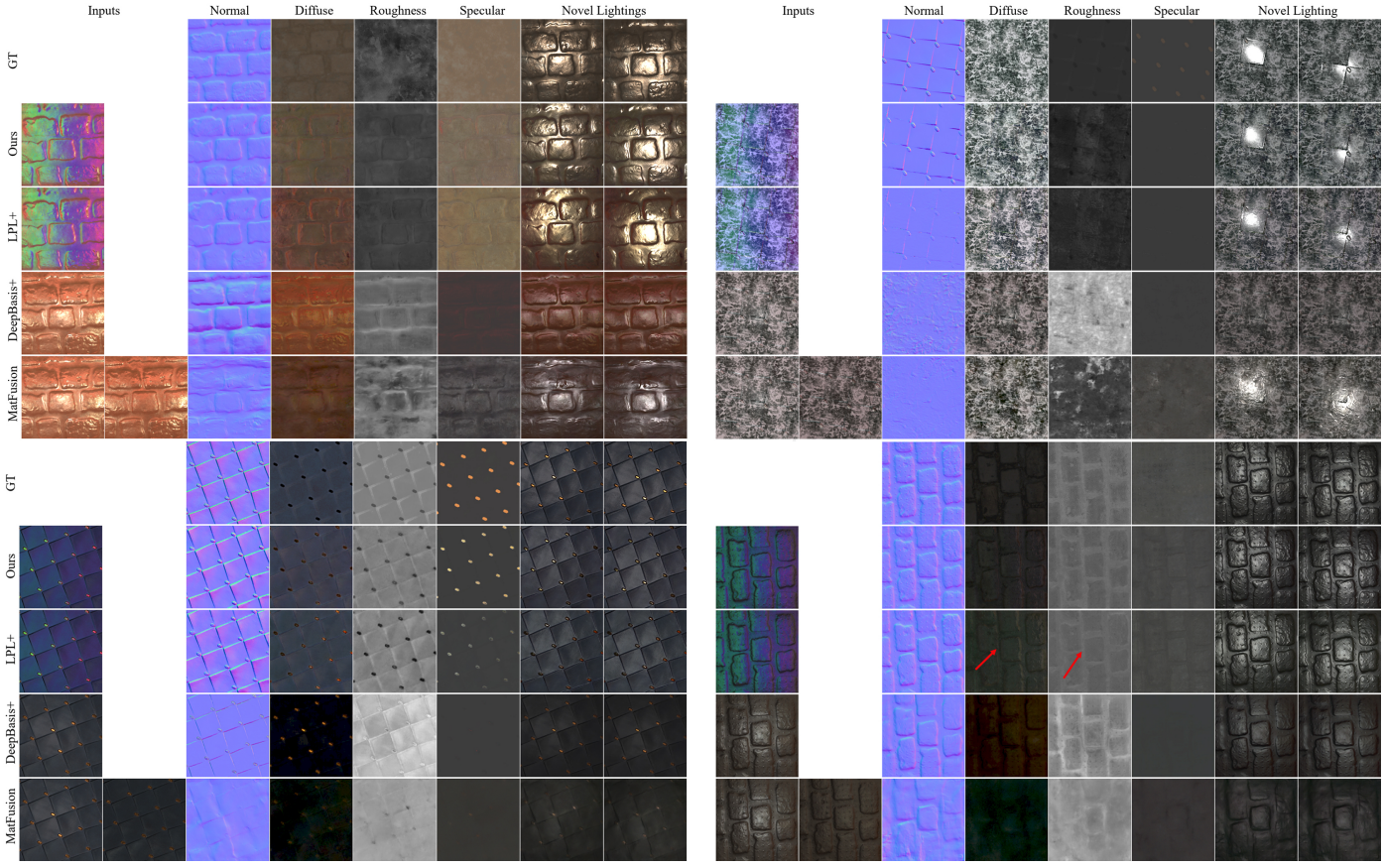


Fig. 10. Comparison against state-of-the-art methods on synthetic data. Here are the reconstruction results of four materials and their re-rendered results under novel point lighting. The red arrows point out the lighting-baked artifacts in LPL+ results.

This proves that our capture setup and proposed lighting clues are effective for resolving ambiguities between uncontrolled lighting and SVBRDFs.

Table 1. Quantitative results of reconstruction results and re-rendering results among our method, previous state-of-the-art methods, and ablation studies. The column abbreviations correspond to "Normals", "Diffuse", "Roughness", "Specular", and "Average", followed by RMSE and LPIPS of the re-rendered results.

Methods	Nrm.	Diff.	Rgh.	Spec.	Avg.	Rend. RSME	Rend. LPIPS
Comparison against SOTA							
DeepBasis+	0.0750	0.0645	0.2431	0.0511	0.1084	0.1535	0.3715
MatFusion	0.0789	0.0575	0.1248	0.0425	0.0759	0.1378	0.3457
LPL+	0.0354	0.0148	0.0322	0.0257	0.0270	0.0732	0.1653
Ours	**0.0277**	**0.0128**	**0.0299**	0.0203	**0.0227**	**0.0588**	**0.1211**
Ablation Studies							
MirrorBall	0.0398	0.0153	0.0337	0.0249	0.0284	0.0746	0.1621
ColoredNLCs	0.0325	0.0147	0.0327	**0.0193**	0.0248	0.0685	0.1402
PlaneLCs	0.0394	0.0162	0.0341	0.0250	0.0287	0.0777	0.1647

The qualitative comparison is visualized in Fig. 10. As shown, our method clearly disentangles the lighting and material. Although LPL+ also uses the same input image as ours captured under active lighting, the environment lighting contributions are highly baked in their reconstructed albedo maps. Besides, DeepBasis+ and MatFusion employ only a point light source, which has limited influence on appearance when captured on-site. This causes the network to be unable to determine whether the material response is caused by environment lighting or the point light source. As a result, significant deviations occur in their reconstruction of reflectance properties. In comparison, although maintaining single-image input, our method can accurately reconstruct various materials. When comparing the results rendered under novel lighting, the renderings produced by our method are also closer to the ground truth (GT) thanks to our introduced mirror ball and semantically aligned lighting clues.

Results on Real Data. To further evaluate our method in real-world environments, we used the acquisition device designed in this study to capture a set of real images and compared the results with other methods. A mobile phone was used as the camera, while an iPad displayed the lighting pattern. Additionally, a 19mm diameter mirror ball was employed to capture the environment lighting and real lighting pattern. The photos captured under this setup were processed and used as input for our method as well as for LPL+. For other methods, the phone's flashlight served as a point light source, and a simple paper frame was used for the calibration of all captured input images. MatFusion utilized two images taken with the phone: one with the flash (Flash) and one without (No-Flash). Deepbasis+ used only the Flash image as input. Besides, the captured environment lighting was in LDR format, which differs from the HDR environment lighting data in training. Thus, we ap-

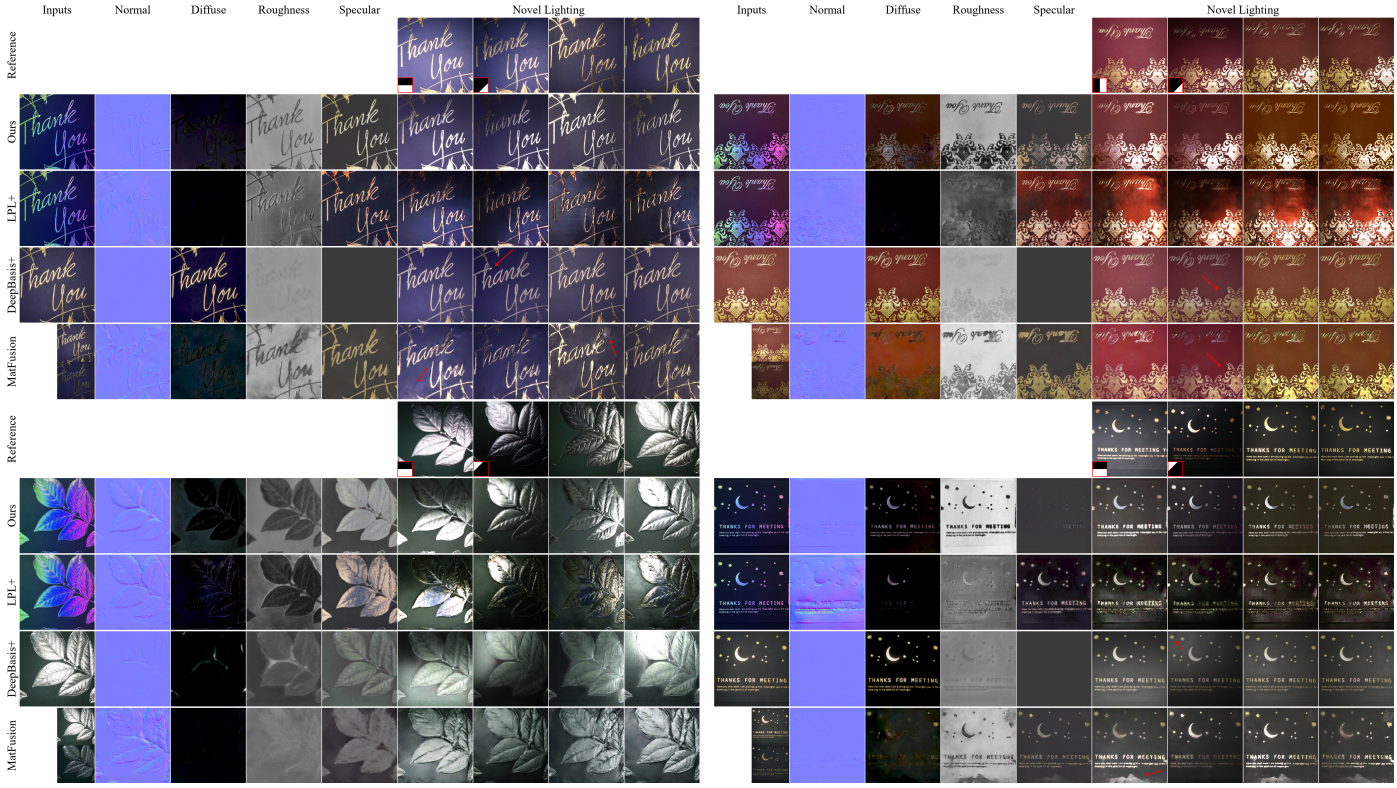


Fig. 11. Comparison against state-of-the-art methods on real data. The first two columns are input images of all methods, while the last three columns show re-render results under novel point lighting for comparison. Images in the first row are reference images for re-renderings, which are captured in a dark room with only a point light source or a planar light source. We mark the step-edge pattern at the left corner of planar lighting-based references. Our predicted normal and roughness are more clean and renderings are more consistent with the ground truth. As pointed by the red arrows, previous methods have artifacts or false highlights.

plied a non-linearity scale to the captured environment lighting for approximation. Given the detected environment lighting and active lighting pattern, our method typically needs about 1.8s to reconstruct the SVBRDFs, where environment prefiltering takes 0.2s, planar lighting clues rendering takes 1.5s, denoise step takes 0.04s and the final reconstruction takes 0.05s. This is higher than LPL+ (0.03s) while lower than DeepBasis+ (3s) and MatFusion (17s).

In Fig. 11, we provide a qualitative comparison of the results from all methods in reconstructing SVBRDFs on four real scenes. Our method demonstrates superior performance across various reflective properties, especially in roughness and specular components. Moreover, we further captured reference images in the darkroom with novel point lighting using cell phone and novel step-edge lighting using LCD screen to evaluate all methods. The latter references are captured under the setup geometry that is roughly aligned with ours. We re-render the reconstructed materials under flash lighting with calibrated flash positions and step-edge lighting with ideal geometry configurations of our setup. Since none of the methods involve color calibration during reconstruction, we adjusted the white balance and intensity of all rendered results to facilitate comparison with the reference images. The reconstructed results of LPL+ are more degraded than those of synthetic data because of the lack of lighting pattern calibration. DeepBasis+ interprets most information in the diffuse component to explain the appearance. Besides, MatFusion exhibits flaws in either the diffuse or roughness reconstruction because of the unstable performance of the generative model. For example, although they correctly reconstructed the relative contrast of roughness levels in the top-right scene, the absolute scale of roughness in the low-roughness regions was overestimated, leading to incorrect highlight rendering results. In contrast, our reconstruction results are more accurate and free of impurities. The rendering results are closer to the references, with reflection effects appearing more realistic.

In addition, instead of conduct comparison under the same environment lighting conditions, we also captured different materials under various environment lighting to validate the robustness of our method. Fig.13 illustrates the SVBRDFs reconstructed by our method under different environment lighting conditions, as well as the results of novel renderings. We capture materials in various real-world environments, such as corridors and halls. It can be observed that our method is robust enough to reconstruct results for on-site capture. In contrast, the SVBRDFs reconstructed by other methods contain more artifacts. The unstable appearance causes blurry texture edges and a loss of material details of their methods. For more results and comparisons, please refer to supplementary materials.

5.2. Ablation Study

Different Clue forms. We evaluated different clue forms to demonstrate the effectiveness of our proposed clue form.

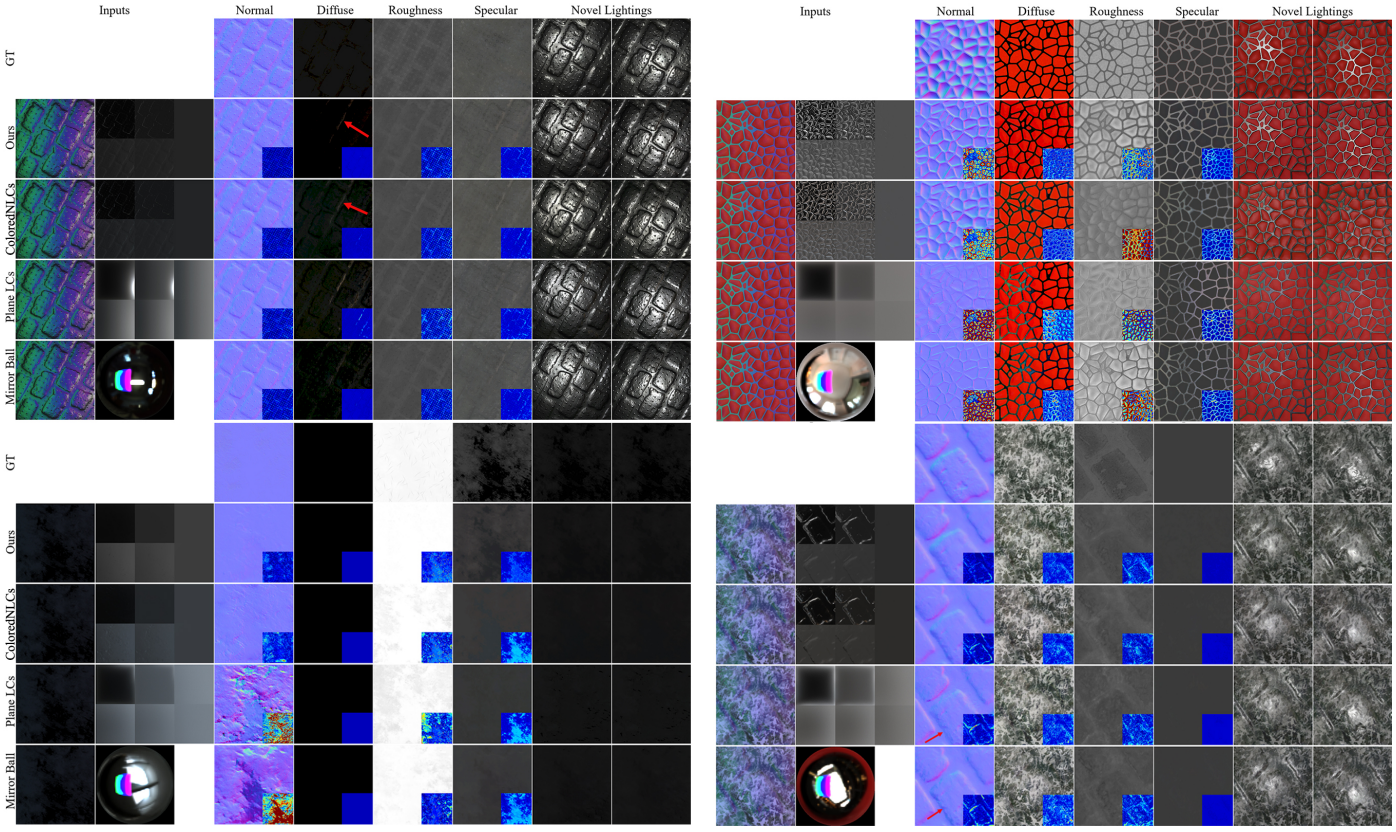


Fig. 12. Qualitative comparison on ablation studies. Input images are shown in the first column, while the different types of lighting clues are shown in the second row. The reconstructed results (3rd-6th column) are re-rendered under two novel point light sources. Red arrows point out the normal artifacts caused by environment lighting.

Specifically, we select to input the mirror ball directly as a comparison. From the quantitative perspective shown in Tab.1, the lighting clues that are semantically aligned with appearance images are proved to be the most effective. In contrast, the SVBRDFs reconstructed using the mirror ball had significantly higher error values compared to our method, especially in normal prediction. This is because the texture structure and semantic information of the mirror ball differ from input images, preventing the network from effectively decoupling active lighting and environment lighting information. As a result, even with clues offered by the mirror ball, it still fails to reconstruct accurate material maps.

Subsequently, a qualitative comparison in Fig. 12 highlights the advantages of the lighting clues in our method: under the mirror ball format, the reconstructed normals exhibit blurry material details and inaccurate edges. The lighting clues used in our method, which share the same semantic information as the appearance map, help the network better understand accurate ambient light information thus leading to the reconstruction of high-quality SVBRDFs. Additionally, the SVBRDFs were re-rendered under novel lighting conditions, and our method produced results closer to GT.

Different types of Lighting Clues. Here we evaluated the impact of normal lighting clues $\hat{I}_i^{lc}$ and plane lighting clues $\bar{I}_i^{lc}$ on reconstructing SVBRDFs. Besides, we also analyzed the impact of color information in $\hat{I}_i^{lc}$. To demonstrate the effectiveness of these three types of lighting clues, we reconstructed the results by inputting different types of lighting clues and retraining our networks. From the quantitative perspective shown in Tab.1, only using $\bar{I}_i^{lc}$ (PlaneLCs) performed poorly in both the reconstruction of reflective properties and the evaluation metrics of the rendered results. This is because the environment lighting information provided by $\bar{I}_i^{lc}$ is inaccurate, failing to capture the influence of environment lighting on the concave and convex parts of the surface. Colored $\hat{I}_i^{lc}$ (ColoredNLCs), on the other hand, showed only a slight advantage in specular property reconstruction but performed poorly in all other attributes. This is primarily due to two reasons: first, when using the split-sum approximation to render Colored $\hat{I}_i^{lc}$, the lighting prefiltering of split-sum introduces color bias in rendering results; second, during the rendering process, Colored $\hat{I}_i^{lc}$ use environment maps with excessive occlusion, and when light sources with color in the map are occluded, color errors will occur. By jointly inputting $\bar{I}_i^{lc}$ and gray-scale $\hat{I}_i^{lc}$, we addressed the issues of inaccurate environment lighting information and color errors, leading to significant improvements in both the quality of the SVBRDFs and the metrics of the re-rendered results.

Through the visualized results in Fig.12, we further conducted qualitative comparisons. For PlaneLCs, the reconstructed normal lost a substantial amount of high-frequency details and edge information, and the roughness exhibited blurred material boundaries. For ColoredNLCs, without gray-scale

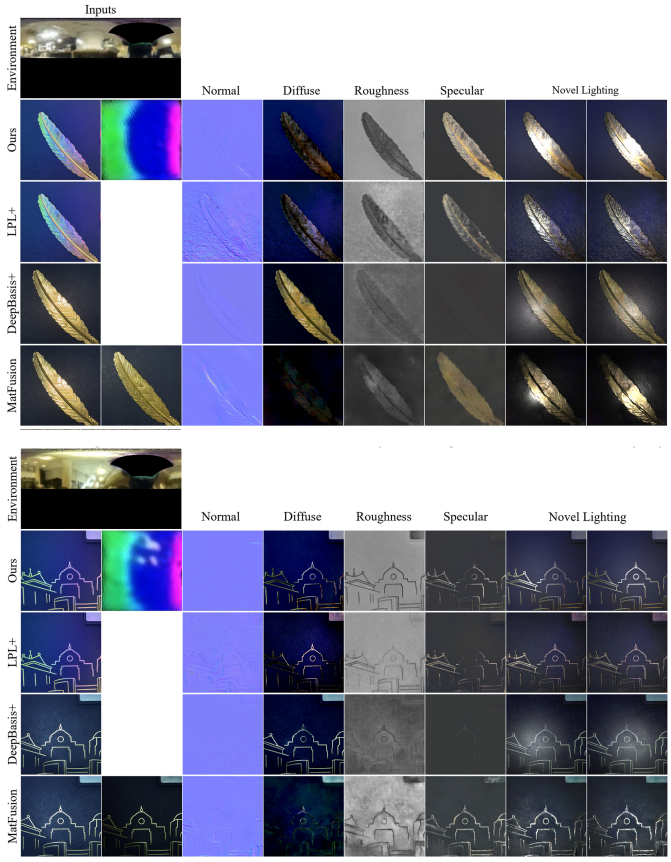


Fig. 13. Comparison against state-of-the-art methods on real data in various environments. The left-top image is environment lighting captured by our mirror ball. The other images in the first two columns are input images and patterns to all methods, respectively. In the last three columns, we also re-render the reconstructed SVBRDF maps (3rd-6th column) under three novel point light sources. Our method robust reconstruct clean SVBRDF maps in different environments.

processing, the reconstructed diffuse was affected by the environment lighting, resulting in noticeable color differences compared to GT, and the roughness contained considerable noise. In contrast, our method reconstructed SVBRDFs with much higher quality, effectively restoring material details. This proves that our combination of $\bar{I}_i^{lc}$ and gray-scale $\hat{I}_i^{lc}$ can effectively hint at environment lighting to the reconstruction network. Furthermore, we conducted rendering experiments under novel lighting conditions, and the comparisons show that our method accurately restored reflection effects.

Reconstruction Robustness on Perturbations. Here, we validate the robustness of our model under different capture perturbations using synthetic data. Specifically, we consider two main types of perturbations for evaluation: geometry perturbations caused by hand-held capture, and pattern display perturbations caused by variations in display devices. For geometry perturbations, we disturb the position and orientation of the light source at different scales. Since our setup relies on relative geometric configurations, this is equivalent to changing the camera geometry. We simulate position perturbations by gradually shifting the lighting rectangle along the x-axis while keep the camera static, with displacement ratios ranging from 2% to 60% of the material size. Note that this distance is usually small in practice thanks to our FOV auto-cropping strategy. For orientation perturbations, we vary the angle between the lighting plane and the xoy-plane, increasing it from 2° to 80°. For pattern display perturbations, we simulate the display process by modifying the lighting pattern with different gamma curves and color balance distortions. Specifically, we apply gamma values ranging from $\frac{1}{2.2}$ to 2.2, and independently scale each color channel by random factors ranging from ±2.5% to ±80%.

In Fig. 15, we report the accuracy curves of our method under each type of capture perturbation. Our method relies on the global correlation between surface points, making it robust to slight geometry perturbations. Although the accuracy decreases as the position and orientation deviate further from the ideal setup, we argue that the relative geometry configuration is simple for common users to achieve. Similarly, thanks to our denoising network, slight pattern display perturbations have minimal impact on reconstruction accuracy.

Table 2. Quantitative results of the denoise network under different pattern settings. RMSE↓ between denoised and ground-truth images is reported.

Patterns	RMSE
Perturbed Ideal Pattern	0.0228
Real Captured Pattern	0.0257

Effectiveness of Denoise Network. Here we conduct two experiments to evaluate the generalization capability of our denoising network on real data. In the first experiment, we capture a real displayed pattern to quantitatively and qualitatively validate the denoise network. Specifically, we render the test set using either a Perturbed Ideal Pattern or a Real Captured Pattern, and then feeding the corresponding images into the denoising network. Then, given these input images and patterns, we measure the RMSE between denoised images and GT images rendered using ideal pattern. The Real Captured Pattern used for rendering is obtained by directly photographing the LCD screen in a darkroom, while the Real Input Pattern for denoise network is captured through the mirror ball under uncontrolled environment. For comparison, the Perturbed Ideal Pattern is generated by applying noise drawn from our training distribution to the Ideal Pattern. The quantitative results are presented in Tab. 2, and the qualitative results are shown in Fig. 16. As demonstrated, even when using real captured patterns for rendering and denoising, our model achieves performance comparable to the synthetic case, which strongly supports its generalization ability to real-world data.

In the second experiment, we further validate the proposed denoise network by comparing the reconstructed SVBRDFs on real data. By inputting the origin captured image and denoised image to the same trained reconstruction network, differences between predicted results are only caused by the denoise network. Besides, to further verify the proposed method for uncontrolled lighting disentangling, we also conduct comparisons against LPL+ given our denoised image as their input. We denote the results directly reconstructed from the origin captured image as w/o Denoise, while denoting the LPL+ equipped with

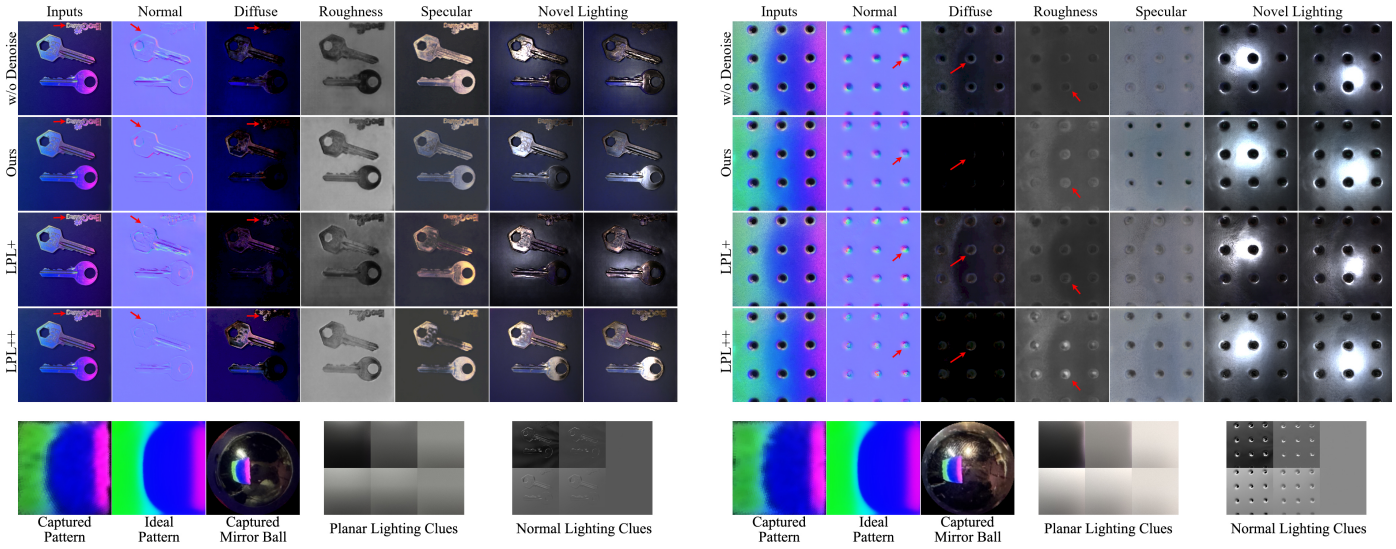


Fig. 14. Validation of denoise network and comparison with LPL+ on real data. The input images of Ours and LPL++ are the same denoised image while others are the same origin captured images. In the last row, we show the real lighting pattern captured by the mirror ball, the ideal pattern for comparison, the mirror ball image to visualize the environment, and proposed lighting clues. Denoise step help both methods improve the accuracy, such as albedo disentanglement and different roughness distinguish. Red arrows in LPL+ and LPL++ results point out the artifacts caused by environment lighting, while our results are more accurate and realistic in comparison.

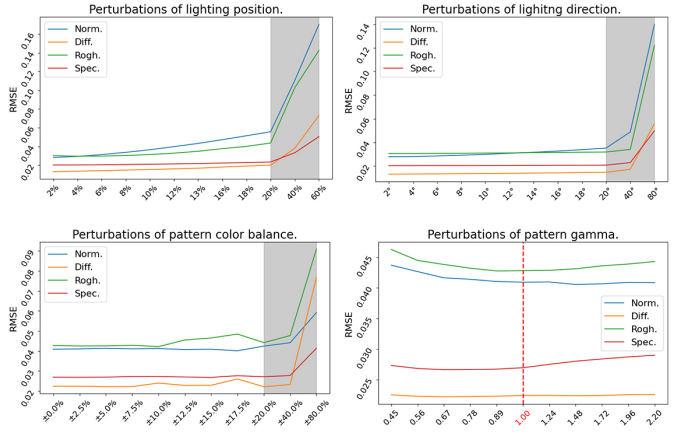


Fig. 15. Robustness evaluation of our method under different capture perturbations on synthetic data. The top row shows the impact of geometry perturbations, including shifts in lighting position (left) and deviations in lighting direction (right). The bottom row shows the impact of pattern display perturbations, including variations in color balance (left) and gamma correction (right). The shaded regions indicate large perturbation ranges that are unlikely in practical setups.

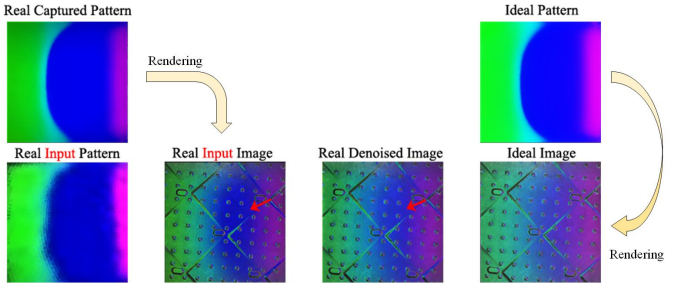


Fig. 16. Qualitative results of our denoise network on real captured pattern. We point out obvious color changing caused by denoising using red arrows.

our denoise network as LPL++.

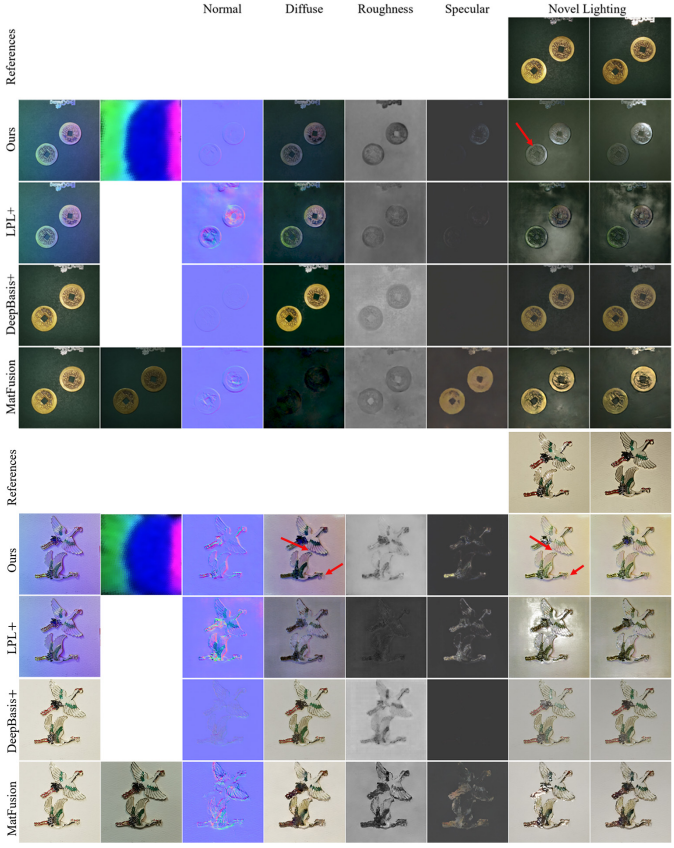
As demonstrated in Fig. 16, although some details are lost because of denoising, our denoised network improves the robustness of both our method and LPL+. Specifically, the roughness of the background card in the left scene and the albedo of metal significantly improved than results predicted from the origin image. In the aspect of comparison, appearance caused by environment lighting is challenging for LPL++ even with denoised input (the last row in Fig. 16). This results in wrong normal and albedo baked-in as pointed out by the red arrows. Contrastingly, thanks to proposed lighting clues that are semantically aligned with captured images, our method can reconstruct accurate normals and albedo maps disentangled with environment lighting. Note that the transition line-like artifacts in the right scene are mainly caused by the difference between the real displayed green color and the ideal pattern. This will be discussed in the limitation.

6. Limitation & Conclusion

6.1. Limitation.

Although our method can reconstruct relatively accurate SVBRDFs under uncontrolled lighting, it still has some limitations. Firstly, as shown in Fig.6.1, the absence of color calibration may cause lighting patterns to be baked into the reconstructed SVBRDF maps, such as the golden coins in the first scene. This is mainly caused by the pattern color difference on real devices from synthetic rendering configuration. While accurate device calibration may solve this problem for our method and LPL+, we argue that the calibration is highly complex for common users. Similarly, shadow caused by partial occlusion of planar lighting to main light source in environment can also

cause baked albedo. This is because that accurate shadow alignment between lighting clues and appearance image can not be guaranteed by rough mirror ball detection. However, thanks to the flexibility of our hand-held capture manner, partial occlusion can be mitigated easily by users though changing capture position. Moreover, when there are significant height variations on the material surface, noticeable self-shadowing effects can occur in areas with complex bumps or depressions. Since our appearance model does not account for such phenomena, the reconstruction accuracy is affected, often resulting in baked-in artifacts in the albedo map.



Limitations of our method on two real scenes. The images in the first two columns are input images and captured patterns to all methods, respectively. The red arrows point to the artifacts caused by no calibration and large height variations.

6.2. *Conclusion*

In this paper, we introduce a lightweight acquisition setup that efficiently captures material properties and utilizes a mirror ball to jointly capture lighting in a single image. Based on this convenient setup, we propose a type of lighting clues to hint at environment lighting contributions in appearance to appearance and thus improve the quality of reconstruction results. Besides, we also design a two-branch network to jointly render normal lighting clues in training and learn data prior for SVBRDF reconstruction. Last but not least, the robustness of our method is significantly improved through a proposed denoise network based on the captured real lighting pattern. Extensive experiments demonstrate that our method outperforms state-of-the-art (SOTA) techniques and performs excellently on a wide range of synthetic and real-world images.

In the future, we aim to combine planar lighting-based material reconstruction with generative modeling. Leveraging planar lighting to stably stimulate reflectance responses in natural scenes, we can use vision model priors for clean and industrial-level material map generation. Besides, inverse rendering of 3D objects guided by extended lighting in uncontrolled environment could be an interesting direction, since 3D objects have large surface variations and thus point lighting can provide limited specularities in SVBRDFs reconstruction.

References

[1] Ghosh, A, Achutha, S, Heidrich, W, O'Toole, M. Brdf acquisition with basis illumination. In: IEEE International Conference on Computer Vision. 2007,.

[2] Holroyd, M, Lawrence, J, Zickler, T. A coaxial optical scanner for synchronous acquisition of 3d geometry and surface reflectance. ACM Transactions on Graphics (TOG) 2010;29(4):1–12.

[3] Deschaintre, V, Aittala, M, Durand, F, Drettakis, G, Bousseau, A. Single-image svbrdf capture with a rendering-aware deep network. ACM Transactions on Graphics 2018;37(4CD):128.1–128.15.

[4] Guo, J, Lai, S, Tao, C, Cai, Y, Yan, LQ. Highlight-aware two-stream network for single-image svbrdf acquisition. ACM Transactions on Graphics 2021;40(4):1–14.

[5] Wang, L, Zhang, L, Gao, F, Zhang, J. Deepbasis: Hand-held single-image svbrdf capture via two-level basis material model. In: SIGGRAPH Asia 2023 Conference Papers. SA '23; New York, NY, USA: Association for Computing Machinery. ISBN 9798400703157; 2023,URL: https://doi.org/10.1145/3610548.3618239. doi:10.1145/3610548.3618239.

[6] Gardner, A, Tchou, C, Hawkins, T, of Southern California Institute for Creative Technologies Graphics Laboratory, PDU. Linear light source reflectometry. ACM Transactions on Graphics (TOG) 2003;.

[7] Zhang, L, Gao, F, Wang, L, Yu, M, Cheng, J, Zhang, J. Deep svbrdf estimation from single image under learned planar lighting. In: ACM SIGGRAPH 2023 Conference Proceedings. SIGGRAPH '23; 2023,doi:10.1145/3588432.3591559.

[8] Lin, Y, Peers, P, Ghosh, A. On-site example-based material appearance acquisition. In: Computer graphics forum; vol. 38. Wiley Online Library; 2019, p. 15–25.

[9] Li, X, Dong, Y, Peers, P, Tong, X. Modeling surface appearance from a single photograph using self-augmented convolutional neural networks. Acm Transactions on Graphics 2017;36(4):45.

[10] Martin, R, Roullier, A, Rouffet, R, Kaiser, A, Boubekeur, T. MaterIA: Single Image High-Resolution Material Capture in the Wild. Computer Graphics Forum 2022;41(2):163–17715 pages. doi:10.1111/cgf.14466.

[11] Miika, , Aittala, , Tim, , Weyrich, , Jaakko, , Lehtinen, . Two-shot svbrdf capture for stationary materials. Acm Transactions on Graphics 2015;.

[12] Aittala, M, Aila, T, Lehtinen, J. Reflectance modeling by neural texture synthesis. ACM Transactions on Graphics 2016;35(4):65.1–65.13.

[13] Boss, M, Jampani, V, Kim, K, Lensch, HPA, Kautz, J. Two-shot spatially-varying brdf and shape estimation. ACM Transactions on Graphics 2020;.

[14] Sartor, S, Peers, P. Matfusion: a generative diffusion model for svbrdf capture. In: ACM SIGGRAPH Asia Conference Proceedings. 2023,URL: https://doi.org/10.1145/3610548.3618194.

[15] Vecchio, G, Deschaintre, V. Matsynth: A modern pbr materials dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024, p. 22109–22118.

[16] Guarnera, D, Guarnera, G, Ghosh, A, Denk, C, Glencross, M. Brdf representation and acquisition. Computer Graphics Forum 2016;35:625–650. doi:10.1111/cgf.12867.

[17] Dong, Y. Deep appearance modeling: A survey. Visual Informatics 2019;3(2):59–68.

[18] Kavoosighafi, B, Hajisharif, S, Miandji, E, Baravdish, G, Cao, W,

Unger, J. Deep svbrdf acquisition and modelling: A survey. In: Computer Graphics Forum; vol. 43. Wiley Online Library; 2024, p. e15199.

[19] Henzler, P, Deschaintre, V, Mitra, NJ, Ritschel, T. Generative modelling of brdf textures from flash images. ACM Transactions on Graphics (TOG) 2021;.

[20] Zhao, Y, Wang, B, Xu, Y, Zeng, Z, Holzschuch, N. Joint svbrdf recovery and synthesis from a single image using an unsupervised generative adversarial network. In: EGSR 2020. 2020,.

[21] Wen, T, Wang, B, Zhang, L, Guo, J, Holzschuch, N. Svbrdf recovery from a single image with highlights using a pre-trained generative adversarial network. In: Computer Graphics Forum. Wiley Online Library; 2022,.

[22] Li, Z, Sunkavalli, K, Chandraker, M. Materials for masses: Svbrdf acquisition with a single mobile phone image. IEEE/CVF Conference on Computer Vision and Pattern Recognition 2018;.

[23] Zhou, X, Kalantari, NK. Look-ahead training with learned reflectance loss for single-image svbrdf estimation. ACM Transactions on Graphics (TOG) 2022;41(6):1–12.

[24] Zhou, X, Kalantari, NK. Adversarial single-image svbrdf estimation with hybrid training. Computer Graphics Forum: Journal of the European Association for Computer Graphics 2021;(2):40.

[25] Vecchio, G, Palazzo, S, Spampinato, C. Surfacenet: Adversarial svbrdf estimation from a single image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021, p. 12840–12848.

[26] Cheng, J, Wang, L, Zhang, L, Gao, F, Zhang, J. Single-image svbrdf estimation with auto-adaptive high-frequency feature extraction. Computers & Graphics 2024;124:104103. URL: https://www.sciencedirect.com/science/article/pii/S0097849324002383. doi:https://doi.org/10.1016/j.cag.2024.104103.

[27] Guo, J, Lai, S, Tu, Q, Tao, C, Zou, C, Guo, Y. Ultra-high resolution svbrdf recovery from a single image. ACM Transactions on Graphics 2023;42(3):1–14.

[28] Nie, Y, Yu, J, Long, C, Zhang, Q, Li, G, Cai, H. Single-image svbrdf estimation using auxiliary renderings as intermediate targets. IEEE Transactions on Visualization and Computer Graphics 2024;.

[29] Luo, X, Scandolo, L, Bousseau, A, Eisemann, E. Single-image svbrdf estimation with learned gradient descent. In: Computer Graphics Forum; vol. 43. Wiley Online Library; 2024, p. e15018.

[30] Luo, D, Sun, H, Ma, L, Yang, J, Wang, B. Correlation-aware encoder-decoder with adapters for svbrdf acquisition. In: SIGGRAPH Asia 2024 Conference Papers. SA '24; New York, NY, USA: Association for Computing Machinery. ISBN 9798400711312; 2024,URL: https://doi.org/10.1145/3680528.3687594. doi:10.1145/3680528.3687594.

[31] Riviere, J, Peers, P, Ghosh, A. Mobile surface reflectometry. In: Computer Graphics Forum; vol. 35. Wiley Online Library; 2016, p. 191–202.

[32] Hwang, I, Jeon, DS, Munoz, A, Gutierrez, D, Tong, X, Kim, MH. Sparse ellipsometry: portable acquisition of polarimetric svbrdf and shape with unstructured flash photography. ACM Transactions on Graphics (TOG) 2022;41(4):1–14.

[33] Baek, SH, Jeon, DS, Tong, X, Kim, MH. Simultaneous acquisition of polarimetric svbrdf and normals. ACM Trans Graph 2018;37(6):268.

[34] Nam, G, Lee, J, Gutiérrez, D, Kim, MH. Practical svbrdf acquisition of 3d objects with unstructured flash photography. ACM Transactions on Graphics (TOG) 2018;.

[35] 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 2019;38(4). URL: https://doi.org/10.1145/3306346.3323042. doi:10.1145/3306346.3323042.

[36] Guo, Y, Smith, C, Hašan, M, Sunkavalli, K, Zhao, S. Materialgan: reflectance capture using a generative svbrdf model. ACM Transactions on Graphics (TOG) 2020;39(6):1–13.

[37] Deschaintre, V, Aittala, M, Durand, F, Drettakis, G, Bousseau, A. Flexible svbrdf capture with a multi-image deep network. Computer Graphics Forum 2019;38(4).

[38] Zhu, P, Lai, S, Chen, M, Guo, J, Liu, Y, Guo, Y. Svbrdf reconstruction by transferring lighting knowledge. In: Computer Graphics Forum; vol. 42. Wiley Online Library; 2023, p. e14973.

[39] Wang, L, Zhang, L, Gao, F, Kang, Y, Zhang, J. Nfplight: Deep svbrdf estimation via the combination of near and far field point lighting. ACM Transactions on Graphics (TOG) 2024;43(6):1–11.

[40] Xian, C, Li, J, Wu, H, Lin, Z, Li, G. Delving into high-quality svbrdf acquisition: A new setup and method. Computational Visual Media 2024;10(3):523–541.

[41] Ye, W, Li, X, Dong, Y, Peers, P, Tong, X. Single image surface appearance modeling with self-augmented cnns and inexact supervision. Computer Graphics Forum 2018;37(7):201–211.

[42] Riviere, J, Reshetouski, I, Filipi, L, Ghosh, A. Polarization imaging reflectometry in the wild. ACM Transactions on Graphics (TOG) 2017;36(6):1–14.

[43] Li, Z, Xu, Z, Ramamoorthi, R, Sunkavalli, K, Chandraker, M. Learning to reconstruct shape and spatially-varying reflectance from a single image. ACM Transactions on Graphics 2018;37(6):1–11.

[44] Vecchio, G, Martin, R, Roullier, A, Kaiser, A, Rouffet, R, Deschaintre, V, et al. Controlmat: a controlled generative approach to material capture. ACM Transactions on Graphics 2024;43(5):1–17.

[45] Rombach, R, Blattmann, A, Lorenz, D, Esser, P, Ommer, B. High-resolution image synthesis with latent diffusion models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021,arXiv:2112.10752.

[46] Cook, RL, Torrance, KE. A reflectance model for computer graphics. Acm Siggraph Computer Graphics 1981;15(3):307–316.

[47] Walter, B, Marschner, SR, Li, H, Torrance, KE. Microfacet models for refraction through rough surfaces. In: Eurographics Symposium on Rendering Techniques. 2007,.

[48] Karis, B, Games, E. Real shading in unreal engine 4. Proc Physically Based Shading Theory Practice 2013;4(3):1.

[49] Williams, L. Pyramidal parametrics. In: Proceedings of the 10th annual conference on Computer graphics and interactive techniques. 1983, p. 1–11.

[50] Chen, L, Chu, X, Zhang, X, Sun, J. Simple baselines for image restoration. In: Avidan, S, Brostow, G, Cissé, M, Farinella, GM, Hassner, T, editors. Computer Vision – ECCV 2022. Cham: Springer Nature Switzerland. ISBN 978-3-031-20071-7; 2022, p. 17–33.

[51] Jakob, W, Speierer, S, Roussel, N, Vicini, D. Dr. jit: A just-in-time compiler for differentiable rendering. ACM Transactions on Graphics (TOG) 2022;41(4):1–19.

[52] Kingma, DP, Ba, J. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980 2014;.

[53] Zhang, R, Isola, P, Efros, AA, Shechtman, E, Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018, p. 586–595.