DeepBasis: Hand-Held Single-Image SVBRDF Capture via Two-Level Basis Material Model

Li Wang
Tianjin University
China
li_wang@tju.edu.cn

Lianghao Zhang
Tianjin University
China
lianghaozhang@tju.edu.cn

Fangzhou Gao
Tianjin University
China
gaofangzhou@tju.edu.cn

Jiawan Zhang*
Tianjin University
China
jwzhang@tju.edu.cn

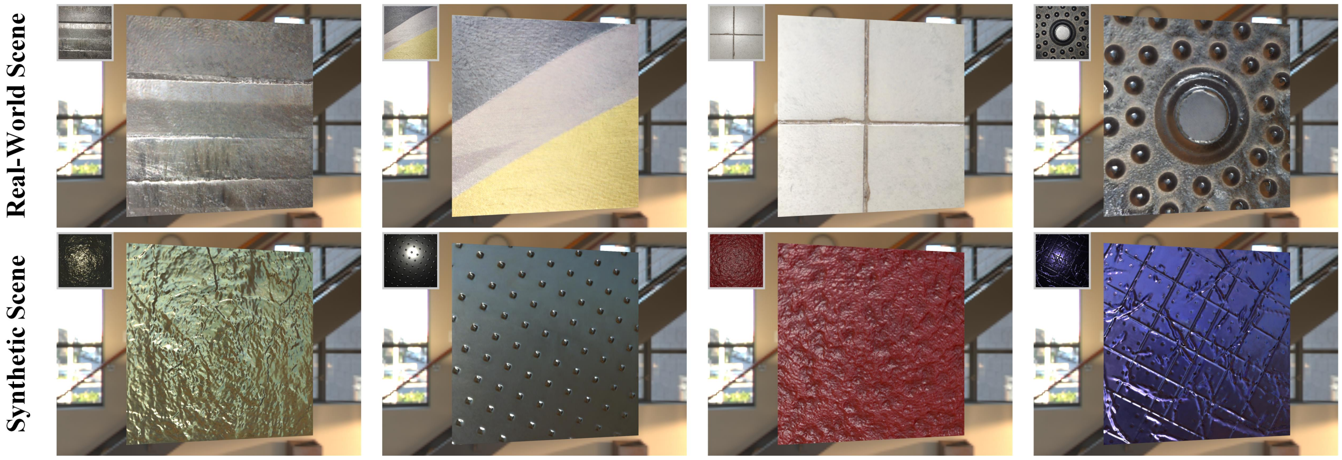


Figure 1: We proposed DeepBasis, a deep-learning-based single-image SVBRDF capture method. Here, we show 4 real-world scenes and 4 synthetic scenes. For each scene, the left-top image is the input image lit by a point light source, and the center visualization is the re-rendering of the estimated material under environment lighting. Thanks to the joint prediction of basis materials and the blending weights, our method can effectively utilize the spatial correlations of materials, thus recovering reflections with rich details.

ABSTRACT

Recovering spatial-varying bi-directional reflectance distribution function (SVBRDF) from a single hand-held captured image has been a meaningful but challenging task in computer graphics. Benefiting from the learned data priors, some previous methods can utilize the potential material correlations between image pixels to serve for SVBRDF estimation. To further reduce the ambiguity from single-image estimation, it is necessary to integrate additional explicit material correlations. Given the flexible expressive ability of basis material assumption, we propose DeepBasis, a deep-learning-based method integrated with this assumption. It jointly predicts basis materials and their blending weights. Then the estimated SVBRDF is their linear combination. To facilitate the extraction of data priors, we introduce a two-level basis model to keep the sufficient representative while using a fixed number of basis materials. Moreover, considering the absence of ground-truth basis materials and weights during network training, we propose a variance-consistency loss and adopt a joint prediction strategy, thereby enabling the existing SVBRDF dataset available for training. Additionally, due to the hand-held capture setting, the exact lighting directions are unknown. We model the lighting direction estimation as a sampling problem and propose an optimization-based algorithm to find the optimal estimation. Quantitative evaluation and qualitative analysis demonstrate that DeepBasis can produce a higher quality SVBRDF estimation than previous methods. All source codes will be publicly released.

*Corresponding authors.

SA Conference Papers '23, December 12–15, 2023, Sydney, NSW, Australia

ACM ISBN 979-8-4007-0315-7/23/12…$15.00
https://doi.org/10.1145/3610548.3618239

CCS CONCEPTS

• **Computing methodologies → Reflectance modeling**.

KEYWORDS

Material Reflectance Modeling, SVBRDF, Basis Maiterials, Deep Learning, Rendering

ACM Reference Format:

Li Wang, Lianghao Zhang, Fangzhou Gao, and Jiawan Zhang. 2023. DeepBasis: Hand-Held Single-Image SVBRDF Capture via Two-Level Basis Material Model. In *SIGGRAPH Asia 2023 Conference Papers (SA Conference Papers '23), December 12–15, 2023, Sydney, NSW, Australia.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3610548.3618239

1 INTRODUCTION

Conveniently estimating the spatial-varying bi-directional reflectance distribution function (SVBRDF) from a single hand-held captured image has been a long-standing problem in computer graphics, with applications ranging from visual product design to virtual/mixed reality and cultural heritage. Such estimation is an inherently ill-posed task because each image pixel value is the only one appearance measurement under its corresponding BRDF, and many different BRDFs can yield the same appearance. Thus, additional information needs to be mined to provide more constraints.

Given the learned prior from data, the potential correlations between image pixels can be utilized to serve for single-image SVBRDF estimation [Li et al. 2017; Ye et al. 2018; Li et al. 2018b,a; Deschaintre et al. 2018; Vecchio et al. 2021; Zhou and Kalantari 2021; Guo et al. 2021; Zhou and Kalantari 2022]. Compared with the implicit correlations learned totally from data, some methods [Zhao et al. 2020; Wen et al. 2022] introduce the self-similarity material assumption [Aittala et al. 2015, 2016] into learning-based methods. It provides an additional explicit constraint above the data priors, thereby further reducing the ambiguity of estimated reflectance maps. However, materials with the self-similar property are very limited, and applying this assumption affects the diversity of captured materials. Therefore, there is a crucial need to pursue a more flexible assumption model that can explicitly establish strong material constraints while adapting to various target materials. Given that data priors are derived from the common features extracted from a substantial amount of data [LeCun et al. 2015; Goodfellow et al. 2016], to integrate with data priors more effectively, this model should possess a consistent and fixed form to facilitate the efficient extraction of common features. Meanwhile, there should be a large amount of relevant training data that can train with this model.

In this paper, we propose DeepBasis, a deep-learning-based approach for single-image SVBRDF estimation. It integrates with the well-known basis material assumption [Matusik et al. 2003b,a; Goldman et al. 2010; Zhou et al. 2016; Kim and Lee 2022], and jointly predicts basis materials and their blending weights. This assumption posits that an SVBRDF can be represented by a set of basis materials, thereby establishing explicit spatial correlations. Meanwhile, the adjustable selection of the number and elemental composition of basis materials provides sufficient flexibility to represent a wide range of materials. Technically, our method includes the following aspects. To facilitate the extraction of data priors, we introduce a two-level model to divide basis materials into global and local components. The adaptive combination of these two components offers the flexible expression capabilities that were previously achieved by altering the basis number. Therefore, even with a fixed basis number, the two-level method still retains a wide representation. To perform training in the absence of ground-truth basis materials and weights, we adopt a joint prediction for both bases and weights. It ensures that SVBRDF can be obtained during each forward pass, thus enabling the utilization of the existing SVBRDF dataset for training. Additionally, to address the potential overlap in the feature extraction of bases and weights during joint training, we introduce a variation-consistency loss as an additional constraint for the predicted basis materials. Moreover, considering the hand-held camera does not guarantee a strict parallel to the material surface, the lighting directions need to be estimated during practical capture. Our observation is that the deviated angle of the camera from being parallel to the material surface is limited. Therefore, we model the estimation of real lighting directions as a sampling problem and propose an optimization-based method to find the optimal estimation.

We evaluated our methods with synthetic and real-world data. The results demonstrated that our DeepBasis could produce better single-image SVBRDF estimation than the existing direct prediction methods [Deschaintre et al. 2018; Zhou and Kalantari 2021, 2022] and the optimization-based method [Gao et al. 2019; Guo et al. 2020].

Overall, our method has the following technical contributions:

- We introduced a two-level basis material model that is specifically designed to fully leverage data priors in the context of basis material assumption.
- We proposed a joint prediction network of basis materials and their blending weights and designed a variation-consistency loss, such that the training process needs no ground-truth basis materials and weights.
- Under the hand-held capture setting, we modeled the real lighting direction estimation as a sampling problem and proposed an optimization-based method to find the optimal real lighting directions.

2 RELATED WORK

Our core idea is to integrate basis material assumption with data priors. Thus, we reviewed deep-learning-based methods and some related methods based on spatial correlation assumptions. Furthermore, considering our adopt near-planar material sample assumption, we focused on reviewing methods with the same assumption. For a more overall discussion of material reflectance acquisition, please refer to the surveys [Guarnera et al. 2016; Dong 2019].

2.1 SVBRDF Estimation Using Deep Learning

2.1.1 Single Input Image. The approaches in this category focus on utilizing material priors learned from data to serve for single-image SVBRDF estimation. Li et al. [2017] proposed a self-augmentation strategy to obtain training data, and Ye et al. [2018] further minimize the requirements for labeled data. Furthermore, some synthetic datasets [Deschaintre et al. 2018; Li et al. 2018a] of SVBRDFs have been proposed, which motivated many SVBRDF reconstruction work based on deep learning [Zhou and Kalantari 2021; Guo et al.

2021; Vecchio et al. 2021; Zhou and Kalantari 2022]. Deschaintre et al.[2020] also proposed a fine-tuning method to capture large-scale planar materials with a few exemplar SVBRDFs. Martin et al. [2022] captured SVBRDF in the wild. Zhou et al. [2022] recently proposed a model to generate tileable SVBRDF. These methods rely on the learned prior from data to implicitly utilize spatial correlations on the target material. In contrast, our method introduced basis material assumption into deep learning, explicitly building the correlations to make deep learning easily extract the required material feature.

2.1.2 Any Number of Input Images. Deschaintre et al. [2019] expanded their single-image work to accommodate any number of input images by introducing an order-independent fusing layer of images. Ye et al. [2021] estimated high-resolution SVBRDF from a flash-lit close-up video sequence captured by a mobile phone. Besides, latent embedded spaces constructed by deep neural networks [Gao et al. 2019; Guo et al. 2020] have also been proposed for SVBRDF optimization. Although these methods could also handle the single-input image, employing multiple images is typically necessary to achieve accurate estimation and avoid overfitting.

2.2 SVBRDF Estimation with Spatial Assumption

2.2.1 Stationary Material Assumption. Wang et al. [2011] proposed that the mesostructure of material surface can be described as a stationary stochastic process. Aittala et al. [2015; 2016] extended the assumption to material reflectance, where similar reflectance properties exist at many different points on the material surface. Based on this, they proposed an optimization-based scheme to reconstruct stochastic SVBRDF using a flash and no-flash image pair.

Leveraging self-similarity assumption, some methods [Zhao et al. 2020; Henzler et al. 2021; Wen et al. 2022] designed an unsupervised generative neural network trained using different small tiles of an input image with similar repetitive features. These methods prove that explicitly defining material correlations in deep learning can effectively improve the prediction accuracy of SVBRDF. However, compared to the basis material assumption, the applied range of the stationary assumption is extremely limited.

2.2.2 Basis Material Assumption. Matusik et al. [2003a; 2003b] proved that any BRDF could be represented by a linear combination of some collected BRDFs. Some methods [Lensch et al. 2003; Chen et al. 2014; Lawrence et al. 2006] extended this observation into SVBRDF estimation and assumed that SVBRDF can be blended by a sparse set of basis BRDFs, which exploits the correlations between different material points. Besides, a pre-collected set of basis materials and sparse blend prior have been employed to reduce the number of input images [Dong et al. 2010; Ren et al. 2011]. Furthermore, a series of methods [Goldman et al. 2010; Zhou et al. 2016; Nam et al. 2018; Alldrin et al. 2008] integrated the pre-collection process of basis materials into an iterative optimization framework along with the blending weights. To improve the effect of optimization, Kim and Lee [2022] proposed a deep embedding clustering-based joint scheme to simultaneously update basis materials and their blending weights. These methods successfully integrate basis material assumption into optimization-based SVBRDF estimation. However, their effectiveness in estimating SVBRDF with sparse measurements, particularly from a single measurement, is constrained by the lack of general priors learned from data. Moreover, the concept of basis materials has also been applied to the representation of Bidirectional Texture Functions (BTF)[Ruiters et al. 2013; Fan et al. 2023].

3 METHOD

3.1 Problem Formulation

Our goal is to estimate spatially-varying material reflectance properties from a single color image. The material sample is assumed to be a nearly planar surface with geometric details that the normal map can model. Similar to Aittala et al. [2015; 2016], we assume that the surface is lit by an approximated point light source co-located with the camera, and the optical axis of the camera is approximately perpendicular to the material sample surface. Note that, we do not need to calibrate the intrinsic and extrinsic parameters of the camera. Moreover, we assume that the reflectance at each material surface point can be well-represented by the Cook-Torrance BRDF model [Cook and Torrance 1981] with GGX microfacet normal distribution [Walter et al. 2007]. Therefore, each SVBRDF can be represented by four material maps: normal map n, diffuse map d, roughness map r, and specular map s. Previous deep learning methods aim to learn a function F to map input image I into its corresponding material maps $M = \{n, d, r, s\}$, as follows:

$$M = F(I). \tag{1}$$

Compared with Eq. 1, introducing basis material assumption, we do not directly predict material maps, but instead a set of basis materials $\{b_i\}$ containing $\{n_i, d_i, r_i, s_i\}$ and their per-point blending weights $\{w_i\}$. Their linear combination is the estimated material maps. Note that, in the standard basis material model used in previous work, b_i is uniform for each surface point. Moreover, we explicitly consider the real lighting directions l under each acquisition.

$$\begin{gathered}\{b_i, w_i\} = F(I, l), \quad i = 1, 2, \cdots, N \\ M = \sum_{i=1}^{N} w_i b_i, \quad w_i, b_i \in [0, 1], \sum_{i=1}^{N} w_i = 1.\end{gathered} \tag{2}$$

To solve Eq. 2, our algorithm needs to meet three requirements. Firstly, it must construct an appropriate basis material model that offers flexibility in adapting to different target materials. Additionally, as the output of F, this model should exhibit a fixed form to facilitate effective training. Secondly, it should find constraints for training without ground-truth $\{b_i, w_i\}$. Finally, it should utilize the single input image I to estimate the real lighting directions l.

3.2 Algorithm

To meet the above requirements, we proposed DeepBasis, as shown in Fig. 2. It takes a single image as the input and predicts basis materials and their blending weights through the network. Then, the estimated SVBRDF is the linear combination of predictions. DeepBasis has three special designs. The first one is a two-level basis material model consisting of global and local components. It

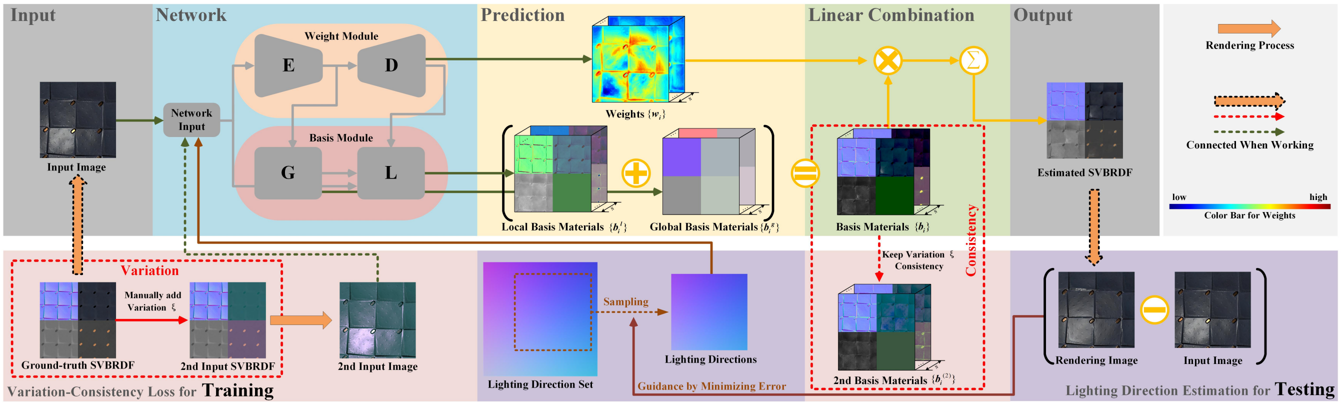


Figure 2: The overview of DeepBasis. It takes a single image as input to the network. The prediction is composed of three elements: weights, local basis materials, and global basis materials. The estimated SVBRDF is the linear combination of weights and the summation of these two basis materials. During the training phase, the introduction of variation-consistency loss ensures that the variation of input SVBRDF is faithfully reflected solely in the predicted basis materials. During the testing phase, the unknown lighting direction can be estimated by iterative sampling within a lighting direction set.

can express a wide range of materials under a fixed form, thereby facilitating the extraction of data priors. The second one is the joint prediction of basis and weight. It makes the network obtain the estimated SVBRDF during each forward pass. Given that the linear combination operation is differentiable, the joint prediction enables the network to be trained with the existing SVBRDF dataset. Besides, we also designed a variation-consistency loss to offer further constraints. The last one is the estimation of real lighting directions, we proposed an optimization-based method to use multiple forward passes of the network to find the optimal lighting directions. Additionally, DeepBasis is well-suited to perform basis refinement after prediction, which makes the estimated SVBRDF closer to the input sample. In the following paragraphs, we discuss the algorithmic details, including two-level basis material model, variation-consistency loss, lighting direction estimation, and basis refinement.

3.2.1 Two-level Basis Material model. To offer enough representation by a small fixed set of basis materials, we proposed that compared with a set of basis materials $\{b_i\}$ shared by the whole material sample in the previous method, each surface point p of the material sample should have its customized basis materials $\{b_i(p)\}$. Thus, the variation of bases on different surface points replaces the changing of basis number, providing sufficient expression capability. Meanwhile, to keep the material correlations for $\{b_i(p)\}$, we proposed a two-level basis material model as follows:

$$b_i(p) = b_i^g + b_i^l(p), \tag{3}$$

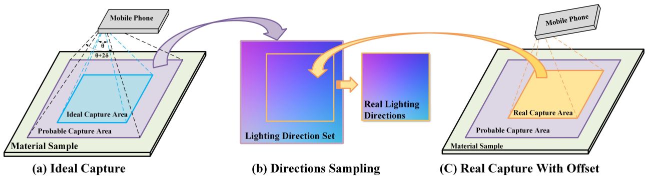
where b_i^g is a global basis material shared by all material surface points and $b_i^l(p)$ is the local one tailored for each surface point. Each $\{b_i(p)\}$ can be regarded as a specific deviation $\{b_i^l(p)\}$ on the global basis material $\{b_i^g\}$. During the back-propagation, the gradients to the global bases have the same optimization directions but those to the local do not. It ensures the global features are more easily accumulated into global bases, thereby offering global material constraints similar to the previous standard basis model, which reduces single-image ambiguity. Meanwhile, the adaptive combination of these two components offers the flexible expression capabilities previously achieved by altering the basis number. Hence, even with a fixed basis number, the two-level model retains broad expressive capacity, making it well-suited for deep learning predictions.

3.2.2 Variation-Consistency Loss. Although the joint prediction strategy makes the existing SVBRDF dataset available for training, relying solely on the SVBRDF constraint leads to ambiguity in the prediction of basis materials and blending weights. To solve it, we design a variation-consistency loss to leverage the variation between twice network forward passes to further constrain the training, as follows:

$$\{b_i^{(1)}, w_i^{(1)}\} = F(R(M), l),\ \{b_i^{(2)}, w_i^{(2)}\} = F(R(M+\xi), l)$$
$$\mathcal{L}_{vc} = ||\{(b_i^{(1)} + \xi) - b_i^{(2)}\}||_1, \tag{4}$$

where R(M) is the rendering of SVBRDF. The loss computation contains twice network forward passes. Through the first one, we can obtain the predicted basis materials $b_i^{(1)}$ and corresponding weights $w_i^{(1)}$ for the input material sample generated by ground-truth SVBRDF M. Before performing the second forward pass, we manually add a random variation ξ into ground-truth SVBRDF to obtain a new SVBRDF $(M + \xi)$ for the generation of the second input sample. Then, through the second pass, we obtain new predicted bases $b_i^{(2)}$ and weights $w_i^{(2)}$. The variation-consistency loss requires the first predicted basis materials plus the variation value equal to the second ones $(b_i^{(1)} + \xi = b_i^{(2)})$, as shown in Fig. 2 (pink box).

The variation ξ has two properties. Firstly, it is uniform for each surface point, ensuring that the spatial structure of the input material remains unchanged. Secondly, it varies across different



(a) Ideal Capture (b) Directions Sampling (C) Real Capture With Offset

Figure 3: Real Capture Modeling. (a) shows the ideal capture, and the generation of the probable capture area. (b) shows the lighting direction set. (c) shows the real capture with camera offset, and the real lighting direction can be seen as a sampling of the lighting direction set.

reflectance maps, enabling the variation-consistency loss to effectively separate the appearance measurements into their respective maps.

3.2.3 Lighting Direction Estimation. As shown in Fig. 3(a) ideal capture, when we assume the maximum offset angle is δ, we can obtain a probable capture area by a FOV $\theta + 2\delta$, where θ is the real camera FOV. In this area, we can calculate each pixel's lighting direction to obtain a set, as shown in Fig. 3(b), and lighting directions of the real capture shown in Fig. 3(c) must be one of its sub-areas. Thus, estimating the real lighting directions under hand-held offset is modeled as finding the optimal sampling from this set. The estimated lighting directions l_e can be computed as follows:

$$l_e = \underset{l_i}{\arg\min} \, ||R(L(F(I, l_i)), l_i) - I||, \tag{5}$$

where l_i represent a possible sampling, and $L(\cdot)$ is the linear combination operation. The optimization process is visually shown in Fig. 2 (bottom purple box). Taking the sampled real lighting directions as network input, we can obtain the estimated SVBRDF, and the rendering image can be computed. We minimize the error between the rendering image and input image to guide finding the optimal real lighting directions. Given that the rendering $R(\cdot)$ is more sensitive to the accuracy of estimated lighting directions than the network prediction $F(\cdot)$, we can speed up the optimization as follows:

$$l_i^{(k+1)} = \underset{l_j}{\arg\min} \, ||R(L(F(I, l_i^{(k)})), l_j) - I||, \tag{6}$$

where $l_i^{(k)}$ represents the k-th l_i of Eq. 5. Given that this equation calculation does not involve the network forward pass, under GPU parallel computing, the taking time is much shorter than performing a network forward pass.

3.2.4 Basis Refinement. Inspired by Gao et al. [2019], we adopt a similar strategy to refine the estimated SVBRDF by minimizing the difference between the input and rendered image. Benefiting from the prediction of basis materials and weights, we can keep the weights fixed and optimize solely global bases.

$$\underset{\{b_i^g\}}{\arg\min} \, ||R(L(\{b_i^g + b_i^l\}, \{w_i\}), l) - I||. \tag{7}$$

The predicted weights possess a well-defined spatial structure. Therefore, optimizing global bases with fewer degrees of freedom effectively prevents local overfitting during single-image optimization. Meanwhile, optimizing global bases enables the utilization of high errors in local regions to refine the global material reflectance. For additional details, please refer to the supplementary materials.

4 IMPLEMENTATION

This section discusses the critical implementation details, including the network architecture, loss function, training details, and testing process. The source code and pre-trained model will be released.

4.1 Network Architecture

As shown in Fig. 2 (top blue box), the network of our DeepBasis consists of weight and basis module. For the weight module, we directly use the architecture proposed by Deschaintre et al. [2018] and modify the input and output. The input has 9 channels, including a 3-channel input image, a 3-channel logarithmic image flattening the dynamic range $[0, 1]$ of the input image, and a 3-channel lighting-mark image. The lighting-mark image is obtained by rendering a fixed SVBRDF using sampled lighting directions. Compared with directly taking directions as input, the rendering makes them unified with other inputs in the image domain. The output is an N-channel weight image and its resolution is the same as the input image.

Given the strong correlation between basis materials and the blending weights, our designed basis and weight modules share the same encoder (E). The basis module comprises two decoding units: a global unit (G) and a local unit (L). The global unit is a multi-layer perceptron (MLP) with 5 layers. Its output is a $N \times 10$ feature map to represent N global basis materials (3-channel normal, 3-channel diffuse, 1-channel roughness, and 3-channel specular) and is tiled to the resolution $N{\times}W{\times}H \times 10$ same as their weights. The local unit estimates the per-pixel basis materials. From the perspective of the resolution, the process from the global basis to the local per-pixel basis can be regarded as a super-resolution (SR) problem. Thus, we directly use a simple but efficient NAFNet architecture proposed by Chu et al. [2022] and Chen et al. [2022] and modify the input and output. Its inputs contain two parts. The first part is 9-channel original inputs, providing the complete information. The second part is the collection of extracted features, paying more attention to the missing features. It includes predicted global basis materials, blending weights, and the error map between the rendering and input image. The output is also a $N{\times}W{\times}H \times 10$ feature map. In our implementation, N is equal to 10, and the effect of the number is evaluated in ablation studies.

4.2 Loss function

Our training loss function has two terms:

$$\mathcal{L} = \lambda_{sup}\mathcal{L}_{sup} + \lambda_{vc}\mathcal{L}_{vc}. \tag{8}$$

The supervised loss $\mathcal{L}_{sup}$ with l1-norm minimizes the error between estimated SVBRDF and ground truth. $\mathcal{L}_{vc}$ is the variation-consistency loss, as discussed in Section 3.2.2, and ξ are sampled on uniform distributions with range $[0, \pi/6]$ for normal vectors and $[0, 0.3]$ for other parameters. These two loss terms are weighted by λ_{sup} and λ_{vc}. For our implementation, $\lambda_{sup} = 1$ and $\lambda_{vc} = 0.05$.

4.3 Training Details

We implemented DeepBasis in PyTorch [Paszke et al. 2019] and trained it with the Adam optimizer [Kingma and Ba 2014] for 400K iterations. The initial learning rate was 2×10^{-5} and was gradually reduced to 1×10^{-7} using the cosine annealing schedule [Loshchilov and Hutter 2016]. The training data were from a public SVBRDF dataset presented by Deschaintre et al. [2018]. The input directions were sampled on a 2-D normal distribution whose mean is the zero offset and whose standard deviation is equal to a third of the maximum offset. In our implementation, the maximum offset angle is $16.33°$. The training takes approximately 12 hours on a single NVIDIA GeForce RTX 4090 graphics card.

4.4 Testing Process

When real lighting directions are pre-known, DeepBasis can directly predict the final results through a network forward pass. For the more common unknown case, DeepBasis first predicts the initial results using ideal lighting directions and then uses the iterative optimization mentioned in Section 3.2.3 to find the optimal estimation and obtain the final prediction results. Benefiting from the acceleration of Eq. 6, it usually takes only a three-times network forward pass to get the final l_e. In our tests, the whole process usually completes within 1 second. When incorporating 500-iterations basis refinement in our implementation, the overall time required for the process can still be maintained within 3 seconds.

5 EXPERIMENTS

In order to evaluate our method, we performed the numerical evaluation and visual analysis on synthetic scenes and real-world captured images and compared the results against the state-of-the-art methods. Moreover, we performed ablation studies to analyze different components' effects on our method.

5.1 Comparison Experiments

We compared our method against the state-of-the-art single-image SVBRDF estimation methods, including RADN [Deschaintre et al. 2018], Hybrid [Zhou and Kalantari 2021] and Look-Ahead [Zhou and Kalantari 2022]. Additionally, we also compared results with the optimization-based methods DIR [Gao et al. 2019] and MGan [Guo et al. 2020]. We obtained their estimation results using the source code and pre-trained models provided by the authors. Note that for DIR, MaterialGan, and Look-Ahead, we provided the known lighting directions as additional inputs.

5.1.1 Comparison on Synthetic Data. We first performed a numerical comparison on a set of 122 synthetic scenes gathered from Deschaintre et al. [2018; 2019], and the results are shown in the upper part of Table 1. Note that these test scenes were never involved in the training. We evaluated the estimated quality of reflectance parameters using Root Mean Square Error (RMSE) and evaluated the re-rendering images using both RMSE and learned perceptual image path similarity (LPIPS) [Zhang et al. 2018]. We performed the re-renderings on 30 random lighting and viewing directions. Our DeepBasis outperformed other methods in estimated material maps and re-rendering quality, as indicated by the RMSE and LPIPS metrics. Furthermore, using basis refinement (Ours+BF) can make the estimated materials produce higher-quality re-rendering appearances.

Table 1: Numerical comparison on 122 synthetic scenes. We evaluate the quality of estimated normal, diffuse, roughness, and specular (N, D, R, S) in terms of RMSE. The re-renderings (Ren.) for each SVBRDF are performed on 30 random lighting directions and evaluated by both RMSE and LPIPS. The lowest errors are highlighted in bold. The upper part is the comparison with the prior methods, and the lower part is the results of ablation studies.

	RMSE					LPIPS
Methods	N	D	R	S	Ren.	Ren.
RADN	0.067	0.044	0.292	0.075	0.091	0.327
DIR	0.073	0.036	0.252	0.067	0.077	0.178
MGan	0.081	0.043	0.210	0.072	0.087	0.263
Hybrid	0.074	0.033	0.167	0.073	0.095	0.199
lookahead	0.065	0.051	0.202	0.083	0.092	0.233
Ours	**0.053**	0.031	0.162	**0.043**	0.076	0.170
Ours+Opt.	**0.053**	**0.30**	**0.161**	**0.043**	**0.071**	**0.134**
w/o $\mathcal{L}_{vc}$	0.053	0.035	0.188	0.065	0.082	0.234
w/o local	0.076	0.033	**0.137**	0.045	0.078	0.350
w/o global	**0.051**	0.036	0.159	0.060	0.084	0.252
w/o bases	0.053	0.037	0.202	0.063	0.085	0.245
baseline	0.052	**0.030**	0.162	**0.043**	**0.062**	**0.153**

Next, we performed a visual comparison and presented two representative synthetic scenes challenging for single-image SVBRDF estimation in Fig. 6. More scenes are available in the supplementary material. We found that RADN fails to reconstruct specular details, while the results of DIR and MGan suffer from artifacts due to single-image overfitting. For Hybrid, the input lighting is baked into reflectance maps (especially diffuse map), resulting in plausible re-rendering results for the scene dominated by diffuse reflectance but obvious artifacts in the decoupled reflectance maps. In Look-Ahead's results, the lack of normal details impedes the re-rendering quality. In contrast, our results can better decouple reflectance maps while having high-quality re-rendering results. Therefore, both the numerical and visual comparisons prove that our estimated SVBRDF is closer to the ground truth (GT) than other methods.

5.1.2 Comparison on Real Data. To evaluate our method, we collected 58 real scenes and captured 9 images with calibrated lighting for each scene, similar to Guo et al. [2020]. Among the captured images, one serves as the input, while the remaining images are utilized as references to evaluate the re-rendering quality of the estimated SVBRDF. Moreover, we conducted evaluations on real-world datasets of MGan and Look-Ahead. The numerical comparison results are presented in the Table. 2, demonstrating that our method is capable of generating re-rendering results that are closer to the references. In Fig. 7, we show two real scenes for the visual comparison, and more results are available in the supplementary material. The first scene is a metal surface with protrusions. For the methods RADN, DIR, MGan and Hybrid, the lighting information is incorrectly baked in the diffuse map. Although Look-Ahead can remove

Table 2: Numerical comparison on real scenes. These scenes are collected from MGan [Guo et al. 2020], Look-Ahead [Zhou and Kalantari 2022], and our real-world capturing. We calculate the difference between the re-rendered and the reference images by RMSE and LPIPS.

	MGan		**Look-Ahead**		**Ours**	
Methods	RMSE	LPIPS	RMSE	LPIPS	RMSE	LPIPS
RADN	0.126	0.367	0.118	0.359	0.172	0.307
DIR	0.148	0.344	0.108	0.300	0.150	0.245
MGan	0.149	0.385	0.121	0.350	0.142	0.403
Hybrid	0.150	0.323	0.141	0.289	0.157	0.262
Look-Ahead	0.117	0.289	0.095	0.230	0.152	0.239
Ours	0.111	0.263	0.084	0.218	0.132	0.231
Ours+Opt.	**0.109**	**0.254**	**0.083**	**0.204**	**0.130**	**0.215**

the lighting effect on reflectance maps, its estimated normal and roughness maps lack the necessary details to reconstruct the specular reflectance, as indicated by the red box. The second scene is a greeting card adorned with sequins, and these sequins are partly lit in the input image, as shown in the red box. The previous methods fall short in recovering the correct roughness or specular maps for these sequins, thus the re-rendered images lack the corresponding highlight details. Benefiting from the utilization of basis materials, our method effectively leverages material spatial correlations to reduce the ambiguity in the overall reflectance map decomposition. Therefore, in the first scene, our estimation exhibits richer normal details, while in the second scene, we can infer a more complete specular map from partially activated highlight information on the sequins.

5.2 Ablation Studies

We conducted ablation studies performed on 122 synthetic scenes to analyze the effects of different components. To solely focus on evaluating the individual effect of different components, the known lighting directions were provided during the testing phrase.

5.2.1 The Effect of Basis Number. To evaluate the effect of basis number on the two-level basis material model, we respectively trained DeepBasis with different basis numbers and compared the estimated results, as shown in Fig 8. We observed that excluding the special case when the basis number is equal to one, the quality of results has no obvious increasing or decreasing trend as the basis number increases. It demonstrates that our two-level model does not require a specific basis number selection to achieve sufficient representational capability. With a lower basis number, the local bases automatically play a more significant role in compensating for the limited expressive ability of the global bases, and vice versa. Consequently, the flexibility in expressive capability that was originally achieved by changing the basis number is replaced by the proportional variation of local bases in the two-level model. Therefore, besides 1, any other basis number is viable, and we used 10 in our implementation.

5.2.2 The Effect of Two-level Basis Material Model. We conducted experiments using only local bases and only global bases to analyze their individual effects on the two-level model. Additionally, to evaluate the overall impact of bases, we removed all bases while maintaining the proposed two-level structure for per-pixel SVBRDF estimation. The numerical evaluation results are presented in Table. 1, and visual results are displayed in Fig. 10.

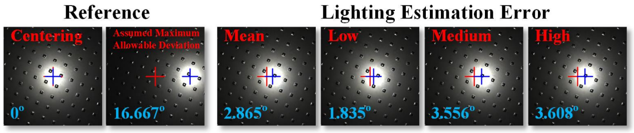


Figure 4: Lighting estimation evaluation. In order to visually represent the angular error, we rendered a synthetic low-roughness material under a point light source. The deviation from the rendered position of a central point light source was used to illustrate the error angle. The left section of the image serves as a visual reference, depicting $0°$ and $16.33°$ angles, while the right section shows the actual representation of estimation errors.

When comparing results without local bases, although the explicit spatial constraint mitigates the adverse effects of over-exposure, it lacks the flexibility to represent fine details. Conversely, without global bases, there is an increase in detail expression; however, it is accompanied by over-exposure-induced weight irregularities and subsequently affects the estimated SVBRDF. However, compared to the results without any bases, the presence of artifacts is still reduced, indicating that per-pixel basis materials still better leverage implicit spatial correlations. These experiments validated that our proposed two-level model (baseline) is reasonable and necessary for the utilization of explicit material correlations and the expression of fine details.

5.2.3 The Effect of Variation-consistency Loss. To evaluate the impact of the variation-consistency loss, we conducted an experiment without it. The results, as shown in Table. 1, indicate degradation in the estimated reflectance maps and re-rendering results. In the context of basis materials, the key avoiding artifacts lies in whether the weights can effectively represent spatial structures. In Fig. 10, we observed partial aliasing in the predictions of basis and weights, leading to the disrupted structures of weights around the central highlight region, consequently affecting the prediction results.

5.2.4 The Effect of Lighting Direction Estimation. To evaluate the accuracy of the estimation, we conducted testing on 10 randomly rendered images per scene under varying lighting directions. The mean error of estimated angular values was calculated across a total of 1220 inputs for evaluation. Considering the significant influence of material roughness on the highlight regions of the images, we categorized the synthetic scenes into low, medium, and high roughness groups. This categorization allowed us to analyze the impact of different roughness levels on lighting estimation, as shown in Fig. 4. Additionally, since the primary objective of lighting estimation is to serve for SVBRDF recovery, we examined the robustness of our method to variations in input lighting directions that deviate from being perpendicular to the sample surface. The evaluation results are presented in Fig. 5. Based on the two experiments, it is

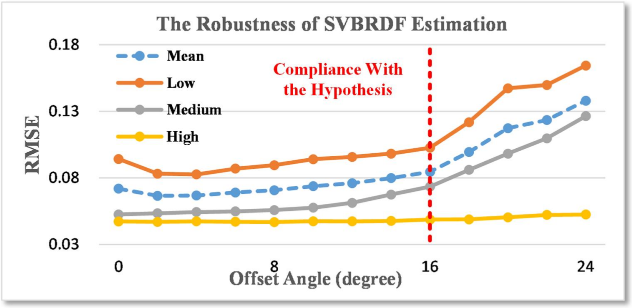


Figure 5: We evaluate the robustness of our method to the input lighting directions using RMSE metrics.

evident that our method accurately estimates the lighting, ensuring stable SVBRDF estimation within the allowed deviation angle (16.33° in our implementation). Furthermore, although materials with high roughness do impact the accuracy of lighting estimation, their SVBRDF estimation exhibits lower sensitivity to variations in input lighting, making the effect negligible.

6 LIMITATIONS AND FUTURE WORK

Single-image SVBRDF estimation is an extremely challenging problem. Although our DeepBasis integrates explicit material spatial correlations into the learned material data priors, further reducing the ambiguity from single-image estimation, we observed that these data priors may still not handle some unconventional materials. Figure 9 offers such an example, where the material sample is a plastic packaging surface with printed 3D patterns. In this case, the printed lighting and shadow patterns might be mistakenly interpreted as reflectance details. In future work, to address this issue, more input images or some strong priors should be provided. For example, if the input sample surface is assumed to be planar with no normal variation, the scene in Fig. 9 may be correctly estimated.

7 CONCLUSION

We have proposed DeepBasis to successfully integrate basis material assumption into learned data priors for single-image SVBRDF estimation. To do so, we proposed a two-level basis material model to ensure the effective extraction of data priors by providing sufficient representation even with a fixed number of bases. Additionally, we adopted the joint prediction method such that the existing SVBRDF dataset can serve for training, and we further designed a variation-consistency loss to avoid the overlap between the feature extractions of bases and weights. Finally, under the hand-held capture setting, we proposed an optimization-based method to estimate the real lighting directions. Extensive experiments on synthetic scenes and real-world captured images demonstrate that our method can produce better results than state-of-the-art methods.

ACKNOWLEDGMENTS

This work was supported in part by National Key Research and Development Program of China (2022YFF0904301).

REFERENCES

Miika Aittala, Timo Aila, and Jaakko Lehtinen. 2016. Reflectance Modeling by Neural Texture Synthesis. *ACM Trans. Graph.* 35, 4, Article 65 (jul 2016), 13 pages. https://doi.org/10.1145/2897824.2925917

Miika Aittala, Tim Weyrich, and Jaakko Lehtinen. 2015. Two-Shot SVBRDF Capture for Stationary Materials. *ACM Trans. Graph.* 34, 4, Article 110 (jul 2015), 13 pages. https://doi.org/10.1145/2766967

Neil Alldrin, Todd Zickler, and David Kriegman. 2008. Photometric stereo with non-parametric and spatially-varying reflectance. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1–8.

Guojun Chen, Yue Dong, Pieter Peers, Jiawan Zhang, and Xin Tong. 2014. Reflectance Scanning: Estimating Shading Frame and BRDF with Generalized Linear Light Sources. *ACM Trans. Graph.* 33, 4, Article 117 (jul 2014), 11 pages. https://doi.org/10.1145/2601097.2601180

Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. 2022. Simple baselines for image restoration. *arXiv preprint arXiv:2204.04676* (2022).

Xiaojie Chu, Liangyu Chen, and Wenqing Yu. 2022. NAFSSR: Stereo Image Super-Resolution Using NAFNet. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 1239–1248.

Robert L. Cook and Kenneth E. Torrance. 1981. A Reflectance Model for Computer Graphics *(SIGGRAPH '81)*. Association for Computing Machinery, New York, NY, USA, 307–316. https://doi.org/10.1145/800224.806819

Valentin Deschaintre, Miika Aittala, Fredo Durand, George Drettakis, and Adrien Bousseau. 2018. Single-Image SVBRDF Capture with a Rendering-Aware Deep Network. *ACM Trans. Graph.* 37, 4, Article 128 (jul 2018), 15 pages. https://doi.org/10.1145/3197517.3201378

Valentin Deschaintre, Miika Aittala, Frédo Durand, George Drettakis, and Adrien Bousseau. 2019. Flexible svbrdf capture with a multi-image deep network. In *Computer graphics forum*, Vol. 38. Wiley Online Library, 1–13.

Valentin Deschaintre, George Drettakis, and Adrien Bousseau. 2020. Guided fine-tuning for large-scale material transfer. In *Computer Graphics Forum*, Vol. 39. Wiley Online Library, 91–105.

Yue Dong. 2019. Deep appearance modeling: A survey. *Visual Informatics* 3, 2 (2019), 59–68.

Yue Dong, Jiaping Wang, Xin Tong, John Snyder, Yanxiang Lan, Moshe Ben-Ezra, and Baining Guo. 2010. Manifold Bootstrapping for SVBRDF Capture. *ACM Trans. Graph.* 29, 4, Article 98 (jul 2010), 10 pages. https://doi.org/10.1145/1778765.1778835

Jiahui Fan, Beibei Wang, Milos Hasan, Jian Yang, and Ling-Qi Yan. 2023. Neural Biplane Representation for BTF Rendering and Acquisition. In *ACM SIGGRAPH 2023 Conference Proceedings*. 1–11.

Duan Gao, Xiao Li, Yue Dong, Pieter Peers, Kun Xu, and Xin Tong. 2019. Deep Inverse Rendering for High-Resolution SVBRDF Estimation from an Arbitrary Number of Images. 38, 4, Article 134 (jul 2019), 15 pages. https://doi.org/10.1145/3306346.3323042

Dan B Goldman, Brian Curless, Aaron Hertzmann, and Steven M. Seitz. 2010. Shape and Spatially-Varying BRDFs from Photometric Stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 6 (2010), 1060–1071. https://doi.org/10.1109/TPAMI.2009.102

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning.* MIT press.

Darya Guarnera, Giuseppe Claudio Guarnera, Abhijeet Ghosh, Cornelia Denk, and Mashhuda Glencross. 2016. BRDF representation and acquisition. In *Computer Graphics Forum*, Vol. 35. Wiley Online Library, 625–650.

Jie Guo, Shuichang Lai, Chengzhi Tao, Yuelong Cai, Lei Wang, Yanwen Guo, and Ling-Qi Yan. 2021. Highlight-Aware Two-Stream Network for Single-Image SVBRDF Acquisition. *ACM Trans. Graph.* 40, 4, Article 123 (jul 2021), 14 pages. https://doi.org/10.1145/3450626.3459854

Yu Guo, Cameron Smith, Miloš Hašan, Kalyan Sunkavalli, and Shuang Zhao. 2020. MaterialGAN: Reflectance Capture Using a Generative SVBRDF Model. 39, 6, Article 254 (nov 2020), 13 pages. https://doi.org/10.1145/3414685.3417779

Philipp Henzler, Valentin Deschaintre, Niloy J. Mitra, and Tobias Ritschel. 2021. Generative Modelling of BRDF Textures from Flash Images. 40, 6, Article 284 (dec 2021), 13 pages. https://doi.org/10.1145/3478513.3480507

Yong Hwi Kim and Kwan H Lee. 2022. Data Driven SVBRDF Estimation Using Deep Embedded Clustering. *Electronics* 11, 19 (2022), 3239.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

Jason Lawrence, Aner Ben-Artzi, Christopher DeCoro, Wojciech Matusik, Hanspeter Pfister, Ravi Ramamoorthi, and Szymon Rusinkiewicz. 2006. Inverse shade trees for non-parametric material representation and editing. *ACM Transactions on Graphics (TOG)* 25, 3 (2006), 735–745.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521, 7553 (2015), 436–444.

Hendrik P. A. Lensch, Jan Kautz, Michael Goesele, Wolfgang Heidrich, and Hans-Peter Seidel. 2003. Image-Based Reconstruction of Spatial Appearance and Geometric Detail. 22, 2 (apr 2003), 234–257. https://doi.org/10.1145/636886.636891

Xiao Li, Yue Dong, Pieter Peers, and Xin Tong. 2017. Modeling Surface Appearance from a Single Photograph Using Self-Augmented Convolutional Neural Networks.

ACM Trans. Graph. 36, 4, Article 45 (jul 2017), 11 pages. https://doi.org/10.1145/3072959.3073641

Zhengqin Li, Kalyan Sunkavalli, and Manmohan Chandraker. 2018a. Materials for Masses: SVBRDF Acquisition with a Single Mobile Phone Image. In *Computer Vision – ECCV 2018*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.). Springer International Publishing, Cham, 74–90.

Zhengqin Li, Zexiang Xu, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. 2018b. Learning to Reconstruct Shape and Spatially-Varying Reflectance from a Single Image. 37, 6, Article 269 (dec 2018), 11 pages. https://doi.org/10.1145/3272127.3275055

Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983* (2016).

Rosalie Martin, Arthur Roullier, Romain Rouffet, Adrien Kaiser, and Tamy Boubekeur. 2022. MaterIA: Single Image High-Resolution Material Capture in the Wild. In *Computer Graphics Forum*, Vol. 41. Wiley Online Library, 163–177.

Wojciech Matusik, Hanspeter Pfister, Matt Brand, and Leonard McMillan. 2003a. A Data-Driven Reflectance Model. *ACM Trans. Graph.* 22, 3 (jul 2003), 759–769. https://doi.org/10.1145/882262.882343

Wojciech Matusik, Hanspeter Pfister, Matthew Brand, and Leonard McMillan. 2003b. Efficient Isotropic BRDF Measurement *(EGRW '03)*. Eurographics Association, Goslar, DEU, 241–247.

Giljoo Nam, Joo Ho Lee, Diego Gutierrez, and Min H Kim. 2018. Practical svbrdf acquisition of 3d objects with unstructured flash photography. *ACM Transactions on Graphics (TOG)* 37, 6 (2018), 1–12.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).

Peiran Ren, Jiaping Wang, John Snyder, Xin Tong, and Baining Guo. 2011. Pocket Reflectometry. 30, 4, Article 45 (jul 2011), 10 pages. https://doi.org/10.1145/2010324.1964940

Roland Ruiters, Christopher Schwartz, and Reinhard Klein. 2013. Example-based Interpolation and Synthesis of Bidirectional Texture Functions. In *Computer Graphics Forum*, Vol. 32. Wiley Online Library, 361–370.

Giuseppe Vecchio, Simone Palazzo, and Concetto Spampinato. 2021. SurfaceNet: Adversarial SVBRDF Estimation from a Single Image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 12840–12848.

Bruce Walter, Stephen R. Marschner, Hongsong Li, and Kenneth E. Torrance. 2007. Microfacet Models for Refraction through Rough Surfaces *(EGSR'07)*. Eurographics Association, Goslar, DEU, 195–206.

Chun-Po Wang, Noah Snavely, and Steve Marschner. 2011. Estimating Dual-Scale Properties of Glossy Surfaces from Step-Edge Lighting *(SA '11)*. Association for Computing Machinery, New York, NY, USA, Article 172, 12 pages. https://doi.org/10.1145/2024156.2024206

Tao Wen, Beibei Wang, Lei Zhang, Jie Guo, and Nicolas Holzschuch. 2022. SVBRDF Recovery from a Single Image with Highlights Using a Pre-trained Generative Adversarial Network. In *Computer Graphics Forum*. Wiley Online Library.

Wenjie Ye, Yue Dong, Pieter Peers, and Baining Guo. 2021. Deep Reflectance Scanning: Recovering Spatially-varying Material Appearance from a Flash-lit Video Sequence. In *Computer Graphics Forum*, Vol. 40. Wiley Online Library, 409–427.

Wenjie Ye, Xiao Li, Yue Dong, Pieter Peers, and Xin Tong. 2018. Single image surface appearance modeling with self-augmented cnns and inexact supervision. In *Computer Graphics Forum*, Vol. 37. Wiley Online Library, 201–211.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yezi Zhao, Beibei Wang, Yanning Xu, Zheng Zeng, Lu Wang, and Nicolas Holzschuch. 2020. Joint SVBRDF Recovery and Synthesis From a Single Image using an Unsupervised Generative Adversarial Network.. In *EGSR (DL)*. 53–66.

Xilong Zhou, Milos Hasan, Valentin Deschaintre, Paul Guerrero, Kalyan Sunkavalli, and Nima Khademi Kalantari. 2022. Tilegen: Tileable, controllable material generation and capture. In *SIGGRAPH Asia 2022 Conference Papers*. 1–9.

Xilong Zhou and Nima Khademi Kalantari. 2021. Adversarial Single-Image SVBRDF Estimation with Hybrid Training. In *Computer Graphics Forum*, Vol. 40. Wiley Online Library, 315–325.

Xilong Zhou and Nima Khademi Kalantari. 2022. Look-Ahead Training with Learned Reflectance Loss for Single-Image SVBRDF Estimation. 41, 6, Article 266 (nov 2022), 12 pages. https://doi.org/10.1145/3550454.3555495

Zhiming Zhou, Guojun Chen, Yue Dong, David Wipf, Yong Yu, John Snyder, and Xin Tong. 2016. Sparse-as-Possible SVBRDF Acquisition. 35, 6, Article 189 (dec 2016), 12 pages. https://doi.org/10.1145/2980179.2980247

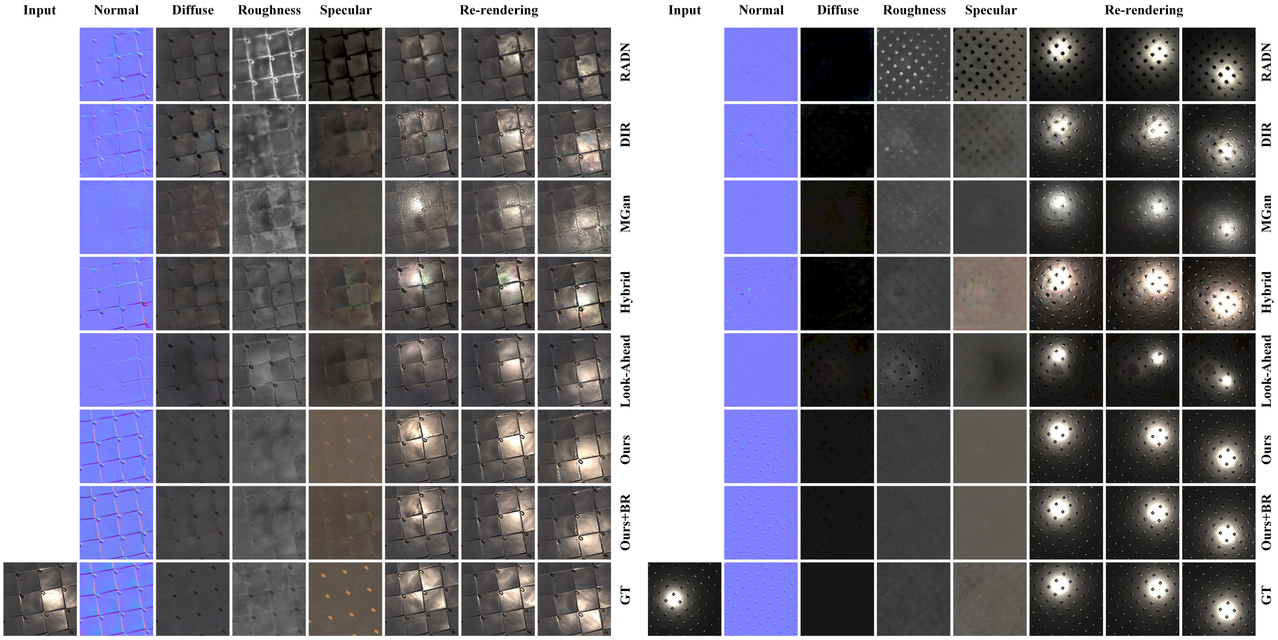


Figure 6: Comparison on Synthetic Data. We compare our results against RADN of Deschaintre et al. [2018], DIR of Gao et al. [2019], MGan of Guo et al. [2020], Hybrid of Zhou et al. [2021] and Look-Ahead of Zhou et al. of [2022] on synthetic data. For each scene, we evaluate the reflectance maps and the re-rendering images using the ground truth (GT).

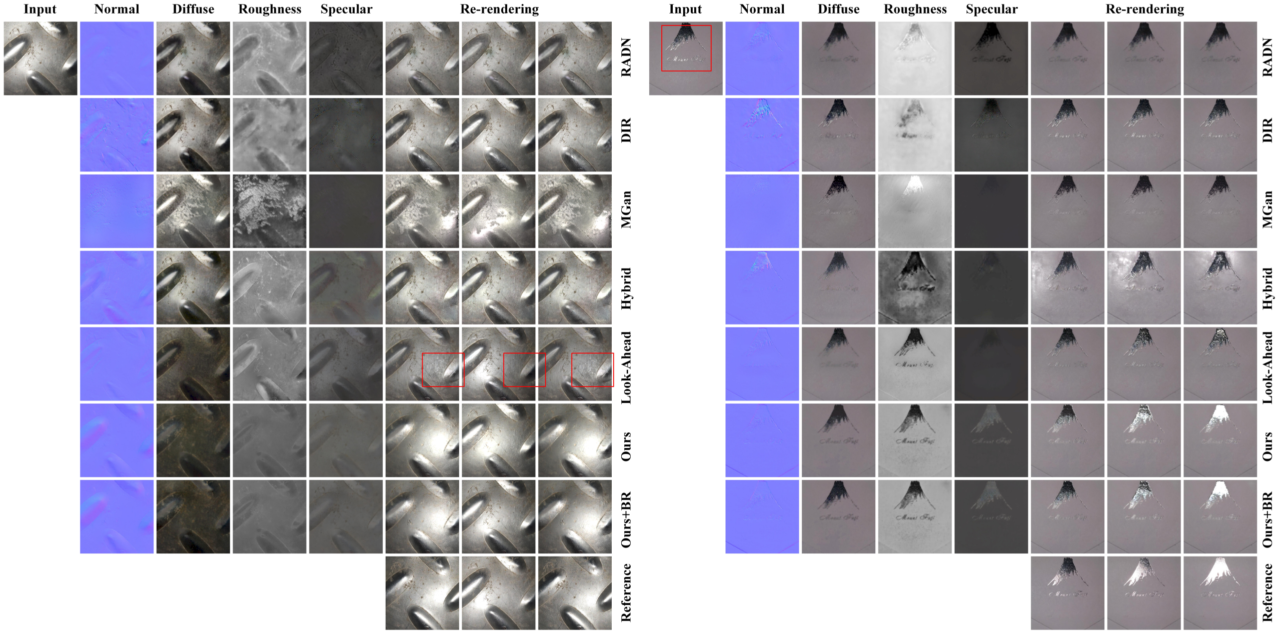


Figure 7: Comparison on Real Data. All input images are captured by a hand-held mobile phone camera with a co-located flashlight. Here, we compare our method against RADN of Deschaintre et al. [2018], DIR of Gao et al. [2019], MGan of Guo et al. [2020], Hybrid of Zhou et al. [2021] and Look-Ahead of Zhou et al. of [2022]. Note that, the required lighting directions of DIR, MGan and Look-Ahead are additionally provided by calibration using a checkerboard.

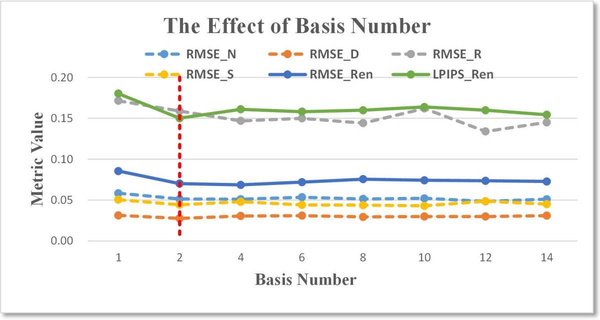


Figure 8: The effect of basis number. The diagram shows the RMSE of SVBRDF and RMSE/LPIPS of re-rendering images under different basis number.

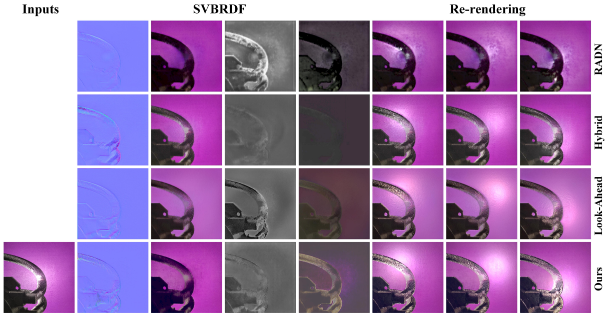


Figure 9: Failure cases. This sample is a plastic packaging surface printed with 3D patterns. The patterns with the lighting and shadows may interfere with the estimation of SVBRDF. We also provide the results estimated by Deschaintre et al. [2018], Zhou and Kalantari [2021; 2022].

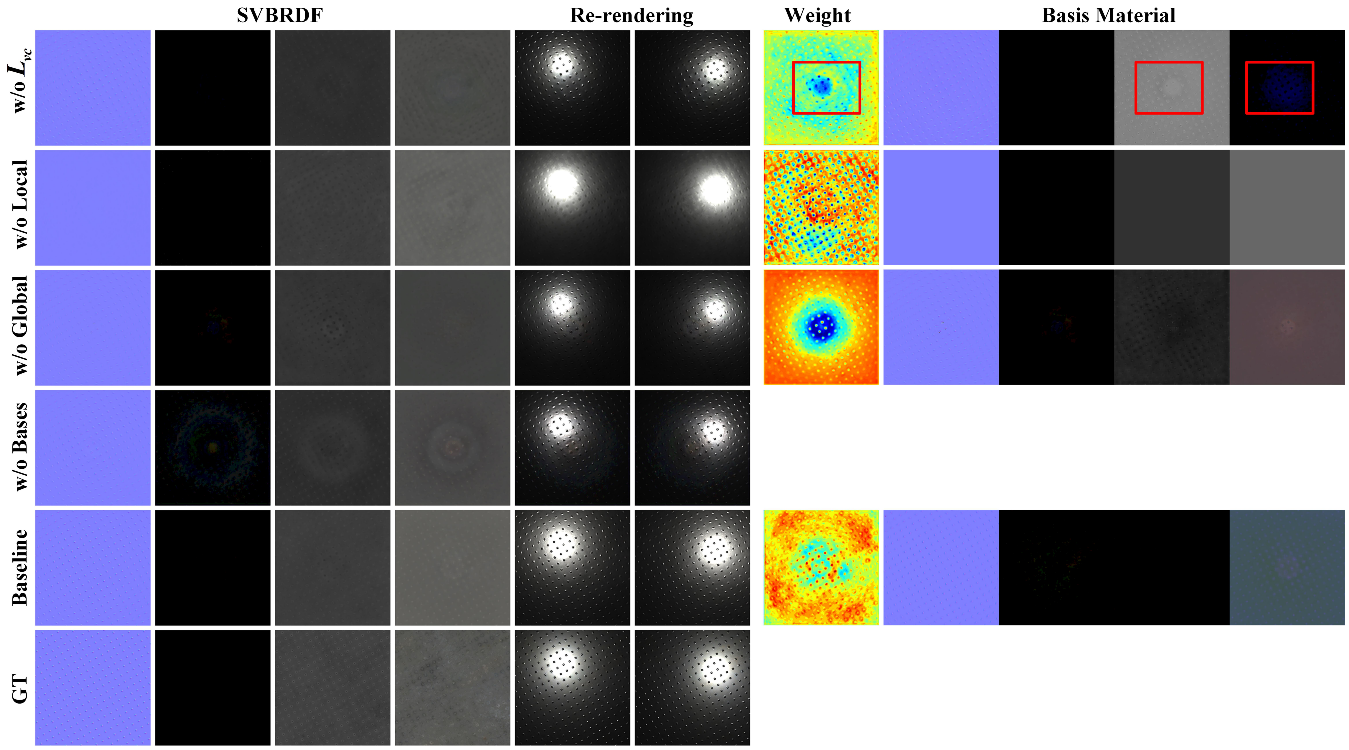


Figure 10: The effect of without different components. In the right part, we show one of basis materials and its corresponding weight to illustrate the direct effect on basis and weight prediction.