Single-Shot Neural Relighting and SVBRDF Estimation

Shen Sang[1,2] and Manmohan Chandraker[1]

[1] University of California, San Diego
[2] ByteDance Research

Abstract. We present a novel physically-motivated deep network for joint shape and material estimation, as well as relighting under novel illumination conditions, using a single image captured by a mobile phone camera. Our physically-based modeling leverages a deep cascaded architecture trained on a large-scale synthetic dataset that consists of complex shapes with microfacet SVBRDF. In contrast to prior works that train rendering layers subsequent to inverse rendering, we propose deep feature sharing and joint training that transfer insights across both tasks, to achieve significant improvements in both reconstruction and relighting. We demonstrate in extensive qualitative and quantitative experiments that our network generalizes very well to real images, achieving high-quality shape and material estimation, as well as image-based relighting. Code, models and data will be publicly released.

Keywords: Single-image relighting, SVBRDF estimation, Physically-based networks

1 Introduction

Single-image relighting is a canonical ill-posed challenge in computer vision, due to the complexity of image formation where spatially-varying material and shape interact with light in myriad ways. Inverse rendering methods have typically recovered shape and material properties, while forward rendering acts on those components for relighting. In this paper, we propose a novel deep network that estimates object shape and material, while jointly learning to relight it under novel illumination conditions, in order to achieve mutual benefits for both tasks. At test time, we use a single image acquired using a flash-enabled commodity mobile phone camera, with possibly unknown environment lighting, to demonstrate recovery of arbitrary object shape and spatially-varying material of complex reflectance, as well as relighting under novel conditions, in a single forward pass.

We achieve this through a physically-motivated modeling and network design. We train a cascaded convolutional neural network (CNN) to estimate shape (depth and surface normals), a spatially-varying bidirectional reflectance distribution function (SVBRDF) consisting of diffuse albedo and specular roughness, as well as

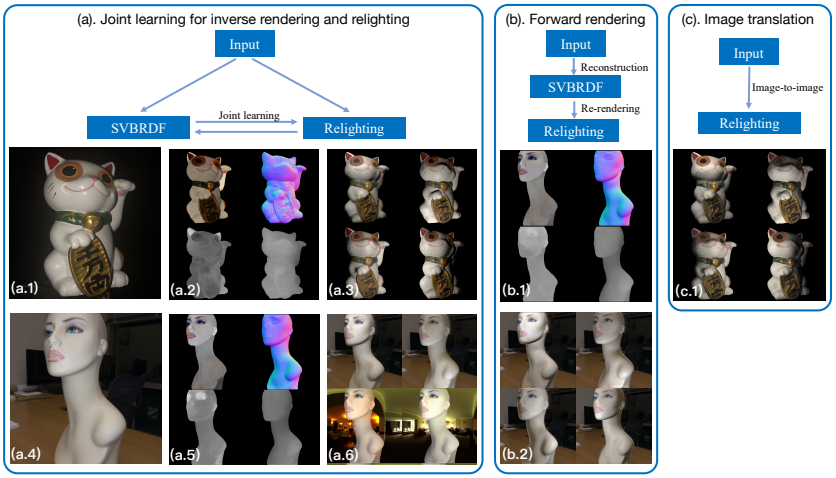

Fig. 1: Three approaches for image relighting with single input images (a.1) and (a.4). **(a)** Our proposed joint inverse rendering and relighting method achieves significant improvements for both tasks, shown in (a.2), (a.3), (a.5) and (a.6). In addition point lights, our method also edits the environment illumination (a.6) **(b)** Forward rendering [19] on recovered shape and material parameters may introduce artifacts (the neck in (b.2)) in the reconstruction stage. **(c)** Image-to-image translation such as [13] lacks physically-motivated modeling and cannot create realistic specularities in (c.1) compared to ours (a.3). More comparisons with forward rendering are shown in Figure 8.

synthesize images under different lighting conditions. Unlike prior works that use a shared encoder but separate decoders for various SVBRDF components [18, 19], we encourage coherence among them through a shared decoder. More importantly, while prior works have considered relighting as a separate forward problem or used "thin" in-network rendering layers that operate at image resolution, we use a deep network that accepts a lighting and decodes the latent shape and SVBRDF codes to an image under novel illumination. This allows us to propose a novel feature sharing mechanism between the inverse rendering and relighting decoders to simulate a physically-based rendering process. Our novel network design, feature sharing and joint training of the inverse and forward tasks allow significant performance improvements in both reconstruction and relighting.

Our reconstruction contrasts with previous works that assume Lambertian or homogeneous material [14], or assume near-planar surfaces as input [7, 17, 18]. It shares insights such as cascaded design with recent works that recover shape and SVBRDF from a single image [19], but goes beyond them in introducing new learning pathways through neural lighting with a shared feature space. For relighting, prior works use multiple images for an exhaustive acquisition [6], or interpolate from sparse samples [20, 25, 26]. The recent method of [35] learns to interpolate using five samples captured under pre-defined directional lights. All these methods produce good relighting results, however, they rely on a large

number of images or specialized hardware for acquisition. In contrast, we produce high-quality relighting using a single image captured with a flash-enabled mobile-phone camera under unknown environment lighting, in a single forward pass of a network, which is cheaper in terms of cost and runtime. Examples are shown in Figure 1. All code, models and data are publicly available.[3]

To summarize, we make the following contributions:

- Joint shape and SVBRDF reconstruction, as well as image relighting, from a single mobile phone image, under point light or environment illumination.
- Physically-based network and feature sharing to jointly learn inverse reconstruction and forward relighting tasks.
- Demonstration of mutual benefits on real images through improved reconstruction and relighting with respect to prior state-of-the-art.

2 Related Works

Shape and Reflectance Reconstruction Shape from shading has been explored with calibrated illumination and Lambertian assumption [14], as also with arbitrary shape and reflectance under natural illumination [23]. A few recent methods reconstruct SVBRDF based on near-planar assumption under unknown natural illumination [17] or collocated flash lighting [1, 7, 18]. Barron and Malik [3] pose the reconstruction problem as one of statistical inference and optimize for the most likely explanation of a single image. Recently, Li et al. [19] propose a deep network to recover shape and SVBRDF from a single image with data-driven priors. In contrast to these approaches, we reconstruct high-quality shape and reflectance properties from a single image jointly with relighting constraints, to achieve improvements on both tasks.

Deep Learning for Inverse Rendering In recent years, deep learning-based methods have shown promising results for inverse rendering problems including illumination estimation [9–11], material recognition [4] and estimation [21], reflectance maps extraction [27], surface appearance recovery [17], normal estimation [2] and depth estimation [8]. For shape and reflectance estimation, near-planar assumption is held in some works for simplicity. To reduce the amount of required labeled training data, Li et al. [17] propose to leverage the appearance information embedded in unlabeled images of spatially varying materials to self-augment the training process. Deschaintre et al. [7] and Li et al. [18] train CNNs to regress SVBRDF and surface normal of a near-planar surface from a single image using in-network rendering to provide additional supervision during training. In contrast, our approach learns to reconstruct shape and BRDF parameters and synthesize relighted images jointly, where the relighting constraint can improve the reconstruction by a large margin compared to the in-network rendering layer.

Image-based Relighting Image-based relighting methods offer realistic rendering of images under novel illumination without modeling the scene by directly reconstructing the light transport function and the reflectance field. Previous

[3] `http://cseweb.ucsd.edu/~viscomp/projects/ECCV20NeuralRelighting/`

methods on light transport acquisition use either brute-force [6] or sparse sampling [20, 25, 26]. Debevec et al. [6] proposes a dense fixed-pattern sampling method to render faces under arbitrary changes in lighting and viewing direction based on recorded imagery. The complete 8D reflectance field, which describes the light transport from the incident light field to the outgoing light field, can be simplified to a 4D function with a fixed viewpoint and 2D incident illumination [25] [26]. Recently, neural network-based method [28] leverages a neural network to exploit the non-linear local coherence in the light transport matrix using sparse image samples. Xu et al. [35] do relighting with five image samples captured under pre-defined directional lights using a deep neural network trained on a large, synthetically rendered dataset. These works demonstrate high-quality results by modeling the complex light transport, but require multiple images to achieve this. In contrast, our method can relight images based on a single input image under a collocated light source, without explicitly learning the light transport. In addition to the relighting constraint, our method also leverages BRDF reconstruction for auxiliary learning. This allows our work to pose the two problems as a joint learning problem in a single network, leading to significant improvements.

Cascaded Network Architecture Cascaded models have been effective in different tasks such as human pose estimation [34, 22], face detection [16] and object detection [5]. For example, Newell et al. [22] proposes eight-stacked hourglass networks to do repeated bottom-up, top-down processing with intermediate supervision to improve the performance of the network and produce more accurate part detection. For shape and SVBRDF estimation, Li et al. [19] use a cascade model to refine the the estimation, with rendering error from the previous stage as input to the following stage. Similarly, our model consists of several cascades, whose effectiveness is shown for both SVBRDF estimation and image relighting in our experiments, for example, in Tables 1 and 2.

3 Method

Given a single image of a complex shape captured under a flash light and environment illumination, our method can reconstruct the shape and spatially-varying BRDF, while simultaneously synthesizing new images under novel lighting. We solve this problem by training a cascaded CNN that derives intuitions from a physically-based rendering process. The framework is illustrated in Figure 2.

Preliminaries and notation Our microfacet BRDF model follows [15]. Let A, N, R, D be the diffuse albedo, normal, roughness and depth, respectively. Let l and v be light and view directions and h be the half-vector. Our BRDF model is:

$$f(A, N, R, l) = \frac{A}{\pi} + \frac{\hat{D}(h, R)\hat{F}(V, h)\hat{G}(l, v, h, R)}{4(N \cdot l)(N \cdot v)} \tag{1}$$

where $\hat{D}(h, R)$, $\hat{F}(v, h)$ and $\hat{G}(l, v, h, R)$ are the distribution, Fresnel and geometric terms, respectively. Since we may use point lights for rendering, depth maps are used for computing the attenuation according to the distance from the light source to the surface. Given (1), an intuitive approach to single-image relighting

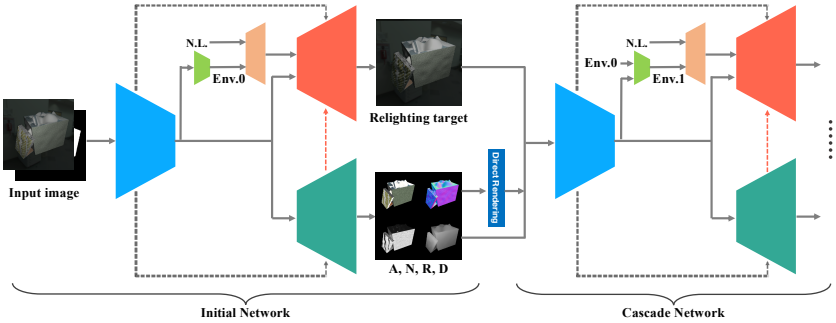

Fig. 2: Overview of the proposed network architecture. We use different colors to visualize various functional components (blue for encoder, green for *InverseDecoder* used for BRDF estimation, and red for *RelightDecoder* used for relighting). Our design consists of an initial model and several cascade stages for iterative refinement. **Left:** The initial model has one encoder for feature extraction and two decoders for BRDF estimation and relighting. The input has either four (without environment lighting) or seven channels (with environment lighting). Besides using skip connections between the encoder and the two decoders, we also feed the features from *InverseDecoder* to *RelightDecoder* to simulate a physically-based rendering process (shown as red dotted arrows). **Right:** The cascade stage is similar to the initial model. It takes the outputs from the first stage and the original inputs as input, leading to a fifteen-channel input. ***Abbreviations:*** N.L: new light. Env.X: the estimated environment map in the X-th cascade. A, N, R, D: albedo, normal, roughness and depth.

is to re-render the image with estimated BRDF parameters. Given estimated parameters $\tilde{A}$, $\tilde{N}$, $\tilde{R}$, with a novel lighting l_{new}, the relighted image I_{new} is:

$$I_{new} = f(\tilde{A}, \tilde{N}, \tilde{R}, l_{new}). \tag{2}$$

3.1 Motivation for Our Design

Although there exist reconstruction methods such as [19] that can produce high-quality estimations, relighting by directly rendering the estimated parameters does not take advantage of any details from the original image. Such image details are removed in the reconstruction step, which results in a loss of detail that might also be useful for new image synthesis. Another intuitive way for relighting can be image-to-image translation [13, 37] where a U-Net [30] architecture is trained, taking a single image and new lighting as input, to generate the relighted target. However, such translation methods fail to create physically reasonable images without knowledge of shape and material. Instead, we seek to bring the best of both worlds, to create relighted images that are more physically meaningful, while preserving details from the input. To this end, we use features from both the original input image and the proposed *InverseDecoder* where shape and material intrinsics are learned. Let f_{enc} and f_{inv} be the two groups of features

extracted by the encoder and the $InverseDecoder$, then in contrast to (2), we propose a $RelightDecoder$ to synthesize an image under new lighting l_{new}:

$$I_{new} = RelightDecoder(f_{enc}, f_{inv}, l_{new}) \tag{3}$$

Jointly training for the forward and inverse tasks has the advantage that our reconstruction effectively benefits from training under different relighting directions, allowing learning of complex relationships between appearance and lighting directions for an object with arbitrary SVBRDF. This arguably allows higher accuracy and better generalization.

3.2 Joint SVBRDF Estimation and Relighting

Our model is built upon an encoder-decoder architecture, which consists of a single shared encoder and two decoders for reconstruction and relighting. Among the three components, there are skip connections used for feature sharing. The input to our encoder consists of a single image, I_{src}, captured under a collocated point light and a mask, M, stacked with the source image. While our formulation is more general, we choose this setup for convenience since a light source collocated with the camera can minimize cast shadows and high-frequency specularities, allowing better observation of the details of shape and material [12, 29].

SVBRDF estimation Unlike [18, 19], which use multiple decoders to reconstruct different parameters, we use only a single decoder, called $InverseDecoder$, to reconstruct the different shape and BRDF parameters: diffuse albedo (A), specular roughness (R), surface normal (N) and depth (D). Since all the parameters correspond to the properties of a particular shape and material, there are internal correlations among them, so we use a single decoder to learn the internal correlations and predict the parameters jointly rather than independently. Compared to [19], we observe that the design for ours not only has faster runtime speed and fewer parameters, but also can achieve higher quality estimations.

Relighting For relighting, we introduce $RelightDecoder$ which takes as input a new lighting vector as well as the feature maps from the encoder and $InverseDecoder$. Instead of being fed into the $RelightDecoder$ directly, the target lighting position, l_{new}, is encoded by a light mapping block that contains three fully connected layers. After concatenating the encoder feature with the lighting vector, we feed it to the $RelightDecoder$, which then creates a new image for the shape and material under a novel light source.

Feature sharing As shown in Figure 2, in addition to using skip connections to transfer the encoder features to the two decoders for retaining spatial details, we also build skip connections between the $InverseDecoder$ and the $RelightDecoder$. This design is inspired by the physically-based rendering where a realistic image is formed by the interaction of shape and BRDF with incident illumination. This allows joint training of the two different but related tasks: reconstruction and relighting, allowing the latent space to encode how appearances vary with light source positions for various shape and SVBRDF configurations. This bidirectional connection provides physical hints for image relighting to produce photorealistic

results, as well as introducing additional supervision for SVBRDF estimation, while most other works tend to use non-learnable in-network rendering layers to achieve this. Thus, this skip connection between leads to significant improvements in relighting and SVBRDF reconstruction compared with previous works.

Standard encoder-decoder methods such as [33, 36] do well on relighting faces with environment lighting, but not for complex shapes with arbitrary SVBRDF under point lights. Qualitatively, we show in Fig 1 that an image translation method does not handle specularities, while our method produces photorealistic outputs since it is physically-motivated. Indeed, the joint learning of relighting and SVBRDF requires new design choices. In Table 1, we quantitatively show that our architecture does better than a single encoder-decoder by ablation study.

Environment map In practice, environment illumination is always present outside of darkroom settings, which has a significant effect on the appearance. In addition to relighting under only a point light, our method can also be generalized to relighting under a point light with an arbitrary unknown environment map. To make our model environment-aware, we append a new branch to our encoder to predict environment maps modeled by spherical harmonics (SH). For each color channel, our network estimates the first nine SH coefficients. Following [19], we add an image with the background to our input which can provide more context information for SH coefficient estimation. Thus, the input to our network with environment map estimation has seven channels.

Cascade refinement The level of detail required for SVBRDF reconstruction and relighting for complex shape is often too high for a single encoder-decoder architecture. Similar to [19], we use a cascaded network to refine our estimation. Let $\tilde{A_n}, \tilde{N_n}, \tilde{R_n}, \tilde{D_n}, I_n^{ren}$ be the SVBRDF estimates and direct rendered image of the n-th cascade. Let $InitialNet$ and $CascadeNet$ be our basic model and cascade models, then our entire model is:

$$\begin{aligned} I_0^{new}, \tilde{A_0}, \tilde{N_0}, \tilde{R_0}, \tilde{D_0} &= InitialNet(I_{src}, M) \\ I_n^{new}, \tilde{A_n}, \tilde{N_n}, \tilde{R_n}, \tilde{D_n} &= CascadeNet(I_{src}, M, \tilde{A_{n-1}}, \\ &\qquad \tilde{N_{n-1}}, \tilde{R_{n-1}}, \tilde{D_{n-1}}, I_{n-1}^{new}, I_{n-1}^{ren}) \end{aligned} \tag{4}$$

The effectiveness of our cascade is quantitatively demonstrated in Table 1 and Table 2. With two cascade stages, both reconstruction and relighting are significantly improved qualitatively and quantitatively.

3.3 Training details

Training data To our best understanding, there is no such large-scale dataset that includes both BRDF parameters and various images illuminated under different light sources. Thus, we adopt the shapes and BRDF parameters of the the synthetic dataset from [19] and render a new dataset. We implement a rendering layer using the BRDF model defined by Equation 1 with PyTorch [24] deep learning framework and CUDA acceleration. Instead of pre-rendering all the training set, we render images in an online manner. For each iteration, we render

relighting target image under random point light sources as the supervision. In this way, our model can see more varied samples as our ground-truth are rendered with a larger set of lights compared with an offline rendering method. We define the position of the point light as the hemisphere in front of the shape.

Network design We use U-Net [30] for $InitialNet$ with large receptive fields to capture the appearance under point light. Our encoder has six convolutional layers with strides of 2. Except for the first layer whose kernel size is 6, all following layers have a kernel size of 4. For $InverseDecoder$, we use transposed convolutions for decoding and skip links to retain details. After deconvolution, we use a residual block and a single convolution layer to yield final SVBRDF parameters. The $RelightDecoder$ consists of six deconvolutional layers, three residual blocks and one output layer. Instead of feeding the feature from the encoder to $RelightDecoder$ directly, we first map the novel lighting to a lighting code using three fully connected layers. The encoded lighting vector is concatenated with the encoder features and then passed to the $RelightDecoder$. We concatenate both the feature from the encoder and $InverseDecoder$ to the $RelightDecoder$. For environment estimation, we pass the highest-level feature from the encoder through two fully connected layers to regress the 3×9 coefficients. For $CascadeNet$, each encoder and decoder contains three convolutional layers and three residual blocks. It has the same feature sharing mechanism among the three modules.

Loss function We use L2 loss as supervision for BRDF estimation and image relighting. We use the inverse transformation in [19] to project it into a fixed range. Given an estimation $\tilde{I}$ and its ground truth I, the L2 loss $\mathcal{L}$ is given by

$$\mathcal{L} = \frac{1}{\Sigma_{i,j} M_{i,j}} \cdot \|(I - \tilde{I}) \cdot M\|_2^2 \tag{5}$$

Let $\mathcal{L}_a$, $\mathcal{L}_n$, $\mathcal{L}_r$, $\mathcal{L}_d$, $\mathcal{L}_{env}$, $\mathcal{L}_{relit}$ be the L2 losses for albedo, normal, roughness, depth, environment map and relighting, the final loss function for our network is:

$$\mathcal{L} = \lambda_a \mathcal{L}_a + \lambda_n \mathcal{L}_n + \lambda_r \mathcal{L}_r + \lambda_d \mathcal{L}_d + \lambda_{env} \mathcal{L}_{env} + \lambda_{relit} \mathcal{L}_{relit} \tag{6}$$

where $\lambda_a = \lambda_n = \lambda_{relit} = 1$, $\lambda_r = \lambda_d = 0.5$ and $\lambda_{env} = 0.01$.

Training strategy We train the networks stage-by-stage. We use a batch size of 16. We use Adam optimizer, with an initial learning rate of 10^{-4} for encoder and 2×10^{-4} for decoders and decrease it by half after every two epochs. We train the initial stage and the two cascade stages for 14, 10 and 9 epochs, respectively.

4 Experiments

We validate the effectiveness of our method with evaluations on synthetic and real data, with comparisons on both relighting and SVBRDF estimation. Please see the supplementary material for several further examples.

4.1 Ablation Study

Feature sharing When training our $RelightDecoder$, we feed the features from the encoder, as well as the features from $InverseDecoder$ to simulate a

	A	*N*	*R*	*D*	Render	Relight	MS-SSIM
Inv	1.64	4.02	4.65	1.80	1.42	-	-
Relit	-	-	-	-	-	1.18	0.867
Inv-Relit-C0	1.57	3.67	4.31	1.74	1.31	1.07	0.888
Inv-Relit-C1	1.43	3.42	4.18	1.55	1.20	1.02	0.889
Inv-Relit-C2	**1.39**	**3.32**	**4.18**	**1.44**	**1.14**	**0.99**	**0.893**

Table 1: Quantitative results of different architecture choices for both SVBRDF reconstruction and relighting under only a point light source. Inv means a *InverseDecoder* is trained. Relit means training a *RelightDecoder*. Inv-Relit means we train the *RelightDecoder* with feature sharing from *InverseDecoder*. The *Render* column in right hand of the table means relighting using the reconstructed BRDFs and the novel lighting. Cn means cascade stage n and C0 is the initial stage. By default, we use L2 error and the magnitude is 10^{-2}. In addition to MSE, we compute the multi-scale structural similarity (MS-SSIM) between our relighting results and its targets.

	I_p^e	I_p^e-bg-C0	I_p^e-bg-C1	I_p^e-bg-C2
Albedo (10^{-2})	1.386	1.324	1.166	**1.157**
Normal (10^{-2})	3.741	3.608	3.344	**3.340**
Roughness (10^{-2})	4.486	4.447	4.305	**4.289**
Depth (10^{-2})	1.802	1.747	1.467	**1.455**
Relight (10^{-3})	9.194	9.062	8.630	**8.626**
Relight (MS-SSIM)	0.892	0.902	0.907	**0.908**

Table 2: Quantitative results for design choices under environment illuminations. I_p^e means an image illuminated by both point light and environment map, -bg means a background image is taken as input.

physics-based rendering process, which gives a better performance compared with that without feature sharing. According to the first three experiments — Inv, Relit, Inv-Relit-C0 in Table 1, it turns out that both the relighting and the reconstruction performance get improved by feature sharing. For relighting, the skip connections between the encoder and *RelightDecoder* give our model the ability to retain details from the original image, while the features from *InverseDecoder* help the *RelightDecoder* to learn a physics-based rendering process. Thus, the relighting performance exceeds both the reconstruction-based method and image-to-image translation method with a large gap. For BRDF reconstruction, a re-rendering loss is proven to be useful in reconstruction tasks [7, 18, 19]. For the *InverseRender*, we apply L2 loss to all the estimated parameters explicitly, as well as an implicit supervision from the relighting branch, which plays a role similar to the re-rendering loss in aforementioned works. In forward phase of training, features from the *InverseDecoder* are passed to the *RelightDecoder*. Then, the gradients from the *RelightDecoder* will be back-propagated to the *InverseDecoder* in backward phase, acting as an implicit supervision. Thus, a joint training of the two task gives a significant improvement for both tasks.

Environment map For the relighting task under an environment mapping as well as a point light, a possible concern can be whether the environment background helps. We train two variants of our *InitialNet* – one with only a masked image as input and the other with both masked and original images as

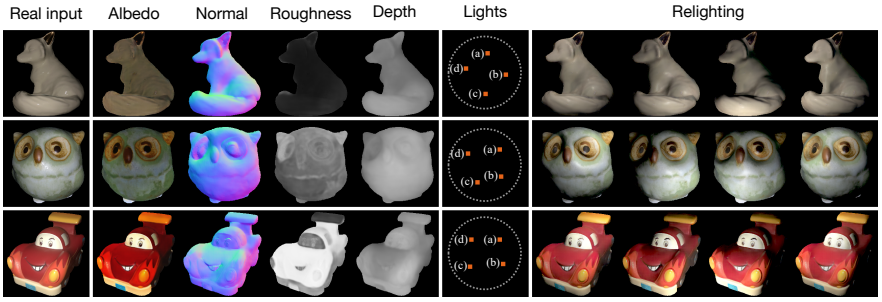

Fig. 3: Reconstruction and relighting results on real data illuminated by the flash light of mobile phone in a darkroom. The input images are shown on the first column, and the following four columns show the BRDF reconstructions. The remaining columns show the light sources and relighting images illuminated under the corresponding lights. The new lights are shown as orange points in a unit circle projected from a hemisphere.

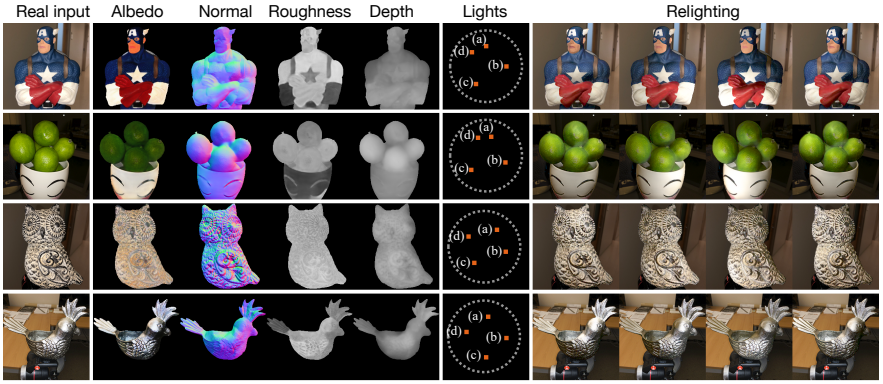

Fig. 4: Results on real images illuminated by a flash light and indoor environment.

input. Quantitative comparison between the first two columns of Table 2 show that both relighting and BRDF estimation are improved with a context input.
Cascade design We show the effectiveness of our cascade design by the quantitative result in Table 1 and 2. By adding two cascades after the *InitialNet*, both of the two tasks are improved significantly. In our experiment, we find that more cascade stages do not provide very significant improvement, and are expensive to train and hard to fit in memory in inference. Thus, two cascade stage suffices.

4.2 Generalization to Real Data

We use real images to demonstrate that our method can generalize to real data in Figure 3 and 4. Real images are captured using an iPhone with the flash enabled. For relighting under a single point light, images are captured in a darkroom. It is

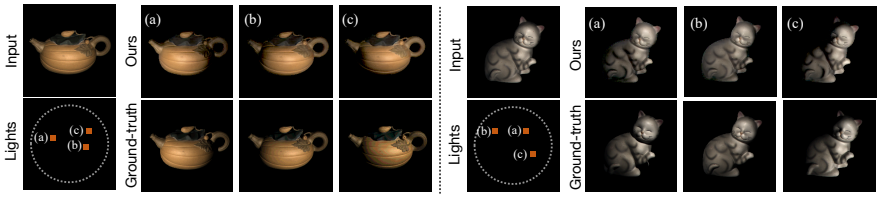

Fig. 5: Relighting results on DiLiGenT dataset consisting of real images. We use the images illuminated by directional lights as reference to demonstrate our relighting performance and the robustness of our method.

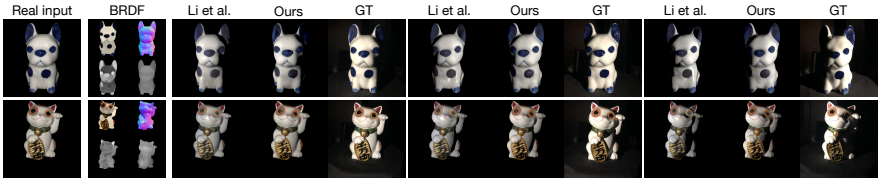

Fig. 6: Comparison between our relighting results with [19], as well as the ground-truth. We capture these ground-truth images using a gantry with a cellphone flash light bound to it. Note that our method can produce realistic results compared to the ground-truth. Limitation: in some cases, cast shadows cannot be produced (e.g. Relit 3 of the cat).

evident that our method produces accurate SVBRDF estimations and relighting outputs, with high-quality shadows and specular highlights under various lights.

We provide ground-truth comparison using two examples in DiLiGenT dataset [31], as well as our own captured images. For DiLiGenT dataset, note that the images are captured under directional lights, so the images are not exactly matched to our relighting task. We demonstrate our approach by using its ground-truth as reference. We take as input the images acquired under a directional light that is approximately collocated to the camera. For each light direction, our method uses a point light source to approximate it. Example results in Figure 5 show that our network is robust enough to yield plausible results on real inputs which do not correspond to the training assumptions. In Figure 6, we show two groups of ground-truth comparison with real images, as well as the results from [19]. Images are captured using a gantry, where we bind a cellphone flash to simulate a point light. The results demonstrate that our method can produce realistic relighting images while also reducing the artifacts.

4.3 Comparative Study

We do comparisons study in two aspects, image relighting and SVBRDF estimation. First, we compare our proposed method with previous works for shape and material estimation and intrinsic image decomposition. Then, we include a comprehensive comparison with [19] to show our superior in both tasks.

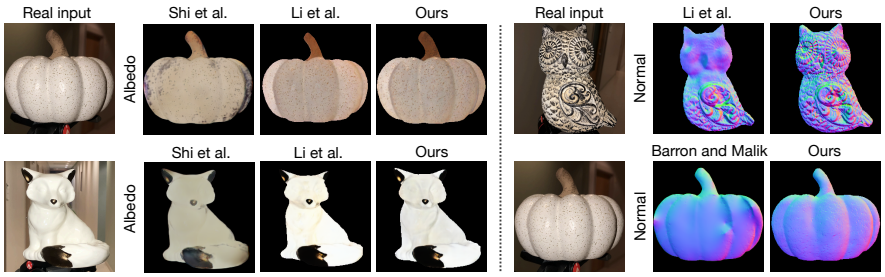

Fig. 7: Comparison on shape and material reconstruction with previous works on real data. **Left:** We compare the diffuse albedo estimation with [32, 19]. Our method outperforms [32] significantly in visual quality and is also comparable with the state-of-the-art method [19]. **Right:** We use two real images to compare the normal estimation with [19, 3]. For a shape full of bumps (the owl), our method produces a more accurate and superior result than [19]. For the pumpkin shape, the accuracy and visual quality of our normal is significantly higher compared with that of [3].

	Albedo	*Normal*	*Rough*	*Depth*
Li et al. [19]	1.215	3.822	4.858	1.505
Ours	**1.157**	**3.340**	**4.289**	**1.455**

Table 3: Quantitative comparison of SVBRDF estimation with [19], using their proposed test set. (MSE, 10^{-2})

Comparisons on SVBRDF estimation with previous works For SVBRDF estimation, we compare our work with [3, 19, 32]. The diffuse albedo estimation with [19, 32] is shown in Figure 7. By comparison, we observe that our SVBRDF estimation is comparable with the state-of-the-art method [19] and outperforms [32] significantly. The result of [32] is smooth and lots of details lost in reconstruction. By contrast, our methods can produce a more detailed estimation. We provide normal estimation comparison in Figure 7. Our method outperforms [19] and [3] and produce more accurate estimations according to the visual quality. Obviously, for the owl shape which is full of bumps, Li et al. [19] fails to recover a high-quality normal, while our method produces a superior result.

Quantitative comparison with [19] for SVBRDF estimation We provide a quantitative comparison with [19] using their test set, in Table 3. We obtain significantly improvements for all components of shape and SVBRDF. This shows that our joint learning allows insights from the forward problem to benefit the inverse problem, in comparison to [19] which focuses only on the inverse problem.

Qualitative comparisons with [19] on real data We also compare with [19] for reconstruction and relighting on real images acquired by a mobile phone camera. To relight in the case of [19], we apply forward rendering using the estimated SVBRDF, while our relighting output is predicted by the network. In Figure 8, green rectangles show artifacts introduced by [19] due to inaccuracy in the estimation of surface normal and roughness, while the visual quality of ours is better. Figure 9 shows that our SVBRDF estimation is also qualitatively better. The video in our supplementary material shows further comparisons.

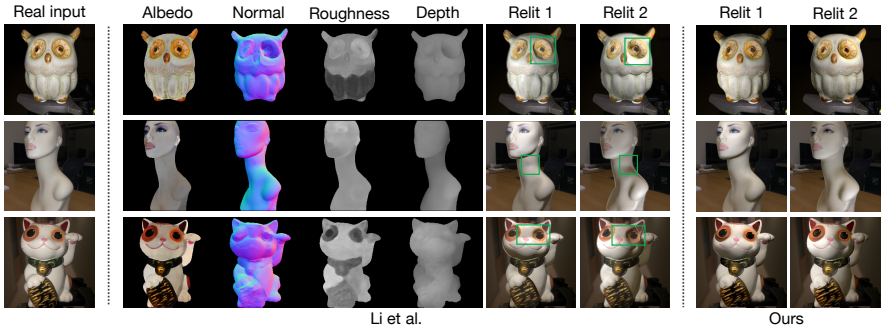

Fig. 8: Comparison with [19] on relighting a real image. Since we learn to jointly relight, our relighting results are less susceptible to errors in shape or SVBRDF estimation.

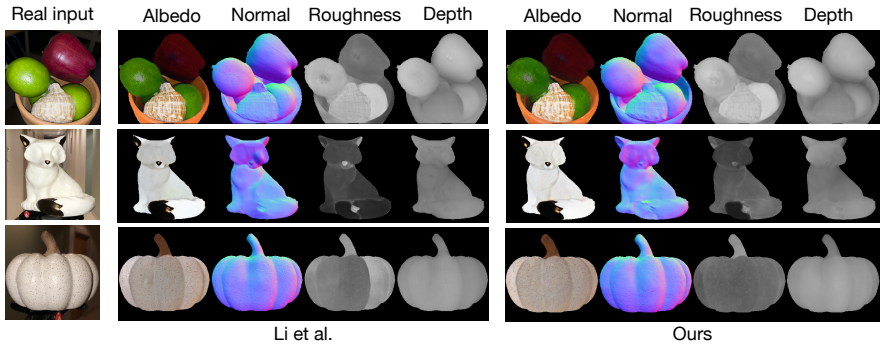

Fig. 9: Comparison with [19] on SVBRDF estimation on real images. While both methods produce high-quality estimates, some factors such as roughness are better estimated by our method.

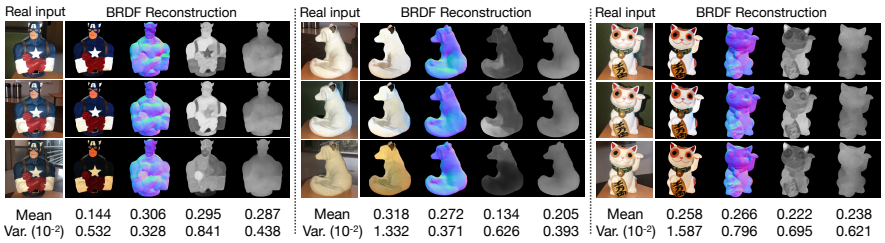

Fig. 10: Mean and variance of SVBRDF estimation under various environment maps.

4.4 Environment Illumination Editing

In addition to relighting under a new light source, our model may also be fine-tuned to allow relighting with a novel environment map. For the new environment

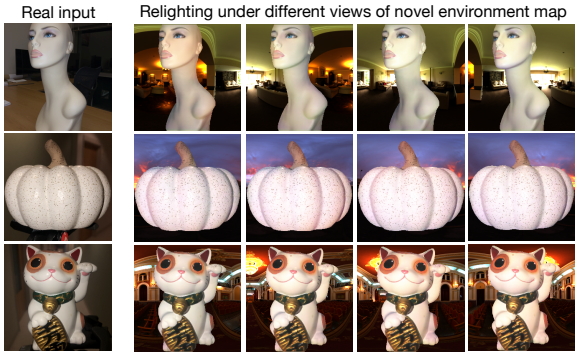

Fig. 11: Given a single real image as input, our network relights it under a new environment map. Four different views are shown in the figure. Note the recovery of fine details, such as the green, violet and red tinges on objects in the three rows, reflecting the colors in the corresponding environment maps.

map, we compute the SH coefficients. Then, we replace the estimated environment coefficients with the new ones as input to the *RelightDecoder*. The environment editing model is fine-tuned on the pre-trained model without any cascade stages involved. Example results are shown in Figure 11. Note that the edited images display realistic shading variations, shadows and specularities. We also include examples of environment illumination editing in the supplementary video.

4.5 Variance under different environment maps

It is non-trivial to obtain ground-truth shape and material measurements aligned with input images. So, we report variance in estimation outputs with different environment maps as an indirect indicator of the accuracy. Using a turntable setup, we acquire images of the same object when illuminated by different parts of an indoor environment. In Figure 10, we show the average across all pixels of the mean and variance of SVBRDF estimates. While some qualitative differences in roughness are understandable for such an underconstrained problem, we observe that variances are quite low, indicating the overall accuracy.

5 Conclusion and Future Work

We present a joint learning approach to reconstruct object shape and SVBRDF, while relighting it under a new light source given only a single image captured by a mobile phone camera. We achieve this by training a cascaded CNN with feature sharing mechanism between the two branches, following the intuition of a physically-based rendering process. Our model is able to progressively refine the estimates, leading to high-quality reconstruction and relighting results on both synthetic and real data. Our future work includes extending the light sources from point lights to a more general illumination. Another direction that can be explored in the future is to study how multiple highly-correlated tasks in forward and inverse rendering can benefit from a joint learning strategy.

Acknowledgments This work was supported by NSF CAREER 1751365, along with generous gifts from a Google Research Award and Adobe Research. This work was done during Shen Sang's graduate studies at UC San Diego.

References

1. Aittala, M., Weyrich, T., Lehtinen, J., et al.: Two-shot svbrdf capture for stationary materials. ACM Trans. Graph. **34**(4), 110–1 (2015)
2. Bansal, A., Russell, B., Gupta, A.: Marr revisited: 2d-3d alignment via surface normal prediction. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5965–5974 (2016)
3. Barron, J.T., Malik, J.: Shape, illumination, and reflectance from shading. IEEE transactions on pattern analysis and machine intelligence **37**(8), 1670–1687 (2014)
4. Bell, S., Upchurch, P., Snavely, N., Bala, K.: Material recognition in the wild with the materials in context database. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3479–3487 (2015)
5. Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6154–6162 (2018)
6. Debevec, P., Hawkins, T., Tchou, C., Duiker, H.P., Sarokin, W., Sagar, M.: Acquiring the reflectance field of a human face. In: Proceedings of the 27th annual conference on Computer graphics and interactive techniques. pp. 145–156. ACM Press/Addison-Wesley Publishing Co. (2000)
7. Deschaintre, V., Aittala, M., Durand, F., Drettakis, G., Bousseau, A.: Single-image svbrdf capture with a rendering-aware deep network. ACM Transactions on Graphics (TOG) **37**(4), 128 (2018)
8. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proceedings of the IEEE international conference on computer vision. pp. 2650–2658 (2015)
9. Gardner, M.A., Sunkavalli, K., Yumer, E., Shen, X., Gambaretto, E., Gagné, C., Lalonde, J.F.: Learning to predict indoor illumination from a single image. arXiv preprint arXiv:1704.00090 (2017)
10. Georgoulis, S., Rematas, K., Ritschel, T., Fritz, M., Tuytelaars, T., Van Gool, L.: What is around the camera? In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5170–5178 (2017)
11. Hold-Geoffroy, Y., Sunkavalli, K., Hadap, S., Gambaretto, E., Lalonde, J.F.: Deep outdoor illumination estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7312–7321 (2017)
12. Hui, Z., Sankaranarayanan, A.C.: Shape and spatially-varying reflectance estimation from virtual exemplars. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(10), 2060–2073 (2016)
13. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. arxiv (2016)
14. Johnson, M.K., Adelson, E.H.: Shape estimation in natural illumination. In: CVPR 2011. pp. 2553–2560. IEEE (2011)
15. Karis, B., Games, E.: Real shading in unreal engine 4. Proc. Physically Based Shading Theory Practice **4** (2013)
16. Li, H., Lin, Z., Shen, X., Brandt, J., Hua, G.: A convolutional neural network cascade for face detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5325–5334 (2015)
17. 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 (TOG) **36**(4), 45 (2017)

18. Li, Z., Sunkavalli, K., Chandraker, M.: Materials for masses: Svbrdf acquisition with a single mobile phone image. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 72–87 (2018)
19. Li, Z., Xu, Z., Ramamoorthi, R., Sunkavalli, K., Chandraker, M.: Learning to reconstruct shape and spatially-varying reflectance from a single image. In: SIGGRAPH Asia 2018 Technical Papers. p. 269. ACM (2018)
20. Matusik, W., Loper, M., Pfister, H.: Progressively-refined reflectance functions from natural illumination. In: Rendering Techniques. pp. 299–308 (2004)
21. Meka, A., Maximov, M., Zollhoefer, M., Chatterjee, A., Seidel, H.P., Richardt, C., Theobalt, C.: Lime: Live intrinsic material estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6315–6324 (2018)
22. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European conference on computer vision. pp. 483–499. Springer (2016)
23. Oxholm, G., Nishino, K.: Shape and reflectance estimation in the wild. IEEE transactions on pattern analysis and machine intelligence **38**(2), 376–389 (2015)
24. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)
25. Peers, P., Dutré, P.: Inferring reflectance functions from wavelet noise. In: Proceedings of the Sixteenth Eurographics conference on Rendering Techniques. pp. 173–182. Eurographics Association (2005)
26. Peers, P., Mahajan, D.K., Lamond, B., Ghosh, A., Matusik, W., Ramamoorthi, R., Debevec, P.: Compressive light transport sensing. ACM Transactions on Graphics (TOG) **28**(1), 3 (2009)
27. Rematas, K., Ritschel, T., Fritz, M., Gavves, E., Tuytelaars, T.: Deep reflectance maps. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4508–4516 (2016)
28. Ren, P., Dong, Y., Lin, S., Tong, X., Guo, B.: Image based relighting using neural networks. ACM Transactions on Graphics (TOG) **34**(4), 111 (2015)
29. Riviere, J., Peers, P., Ghosh, A.: Mobile surface reflectometry. In: Computer Graphics Forum. vol. 35, pp. 191–202. Wiley Online Library (2016)
30. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
31. Shi, B., Wu, Z., Mo, Z., Duan, D., Yeung, S.K., Tan, P.: A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3707–3716 (2016)
32. Shi, J., Dong, Y., Su, H., Yu, S.X.: Learning non-lambertian object intrinsics across shapenet categories. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1685–1694 (2017)
33. Sun, T., Barron, J.T., Tsai, Y.T., Xu, Z., Yu, X., Fyffe, G., Rhemann, C., Busch, J., Debevec, P.E., Ramamoorthi, R.: Single image portrait relighting. ACM Trans. Graph. **38**(4), 79–1 (2019)
34. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4724–4732 (2016)
35. Xu, Z., Sunkavalli, K., Hadap, S., Ramamoorthi, R.: Deep image-based relighting from optimal sparse samples. ACM Transactions on Graphics (TOG) **37**(4), 126 (2018)

36. Zhou, H., Hadap, S., Sunkavalli, K., Jacobs, D.W.: Deep single-image portrait relighting. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 7194–7202 (2019)
37. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Computer Vision (ICCV), 2017 IEEE International Conference on (2017)