Learning to Reconstruct Shape and Spatially-Varying Reflectance from a Single Image

ZHENGQIN LI, University of California, San Diego
ZEXIANG XU, University of California, San Diego
RAVI RAMAMOORTHI, University of California, San Diego
KALYAN SUNKAVALLI, Adobe Research
MANMOHAN CHANDRAKER, University of California, San Diego

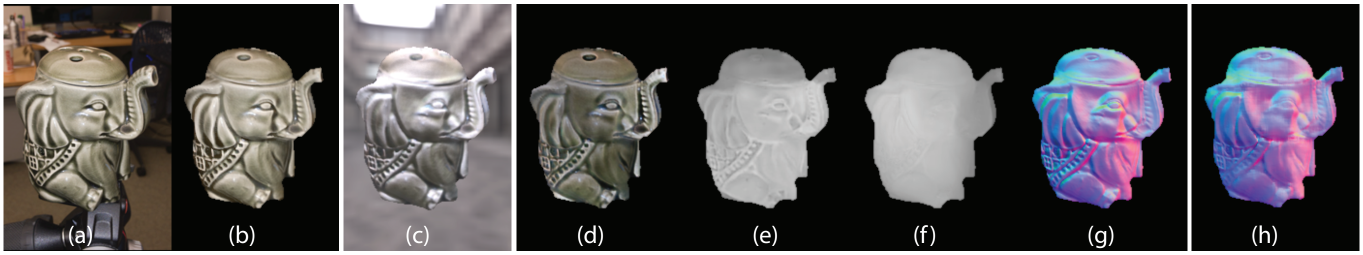


Fig. 1. We propose a novel physically-motivated cascaded CNN architecture for recovering arbitrary shape and spatially-varying BRDF from a single mobile phone image. (a) Input image in unconstrained indoor environment with flash enabled. (b) Relighting output using estimated shape and SVBRDF. (c) Rendering output in novel illumination. (d–g) Diffuse albedo, roughness, depth and surface normals estimated using our framework. (h) Normals estimated using a single-stage network. Our cascade design leads to accurate outputs through global reasoning, iterative refinement and handling of global illumination.

Reconstructing shape and reflectance properties from images is a highly under-constrained problem, and has previously been addressed by using specialized hardware to capture calibrated data or by assuming known (or highly constrained) shape or reflectance. In contrast, we demonstrate that we can recover non-Lambertian, spatially-varying BRDFs and complex geometry belonging to any arbitrary shape class, from a single RGB image captured under a combination of unknown environment illumination and flash lighting. We achieve this by training a deep neural network to regress shape and reflectance from the image. Our network is able to address this problem because of three novel contributions: first, we build a large-scale dataset of procedurally generated shapes and real-world complex SVBRDFs that approximate real world appearance well. Second, single image inverse rendering requires reasoning at multiple scales, and we propose a cascade network structure that allows this in a tractable manner. Finally, we incorporate an in-network rendering layer that aids the reconstruction task by handling global illumination effects that are important for real-world scenes. Together, these contributions allow us to tackle the entire inverse rendering problem in a holistic manner and produce state-of-the-art results on both synthetic and real data.

Authors' addresses: Zhengqin Li, University of California, San Diego, zhl378@eng.ucsd.edu; Zexiang Xu, University of California, San Diego, zexiangxu@cs.ucsd.edu; Ravi Ramamoorthi, University of California, San Diego, ravir@cs.ucsd.edu; Kalyan Sunkavalli, Adobe Research, sunkaval@adobe.com; Manmohan Chandraker, University of California, San Diego, mkchandraker@eng.ucsd.edu.

0730-0301/2018/11-ART269 $15.00
https://doi.org/10.1145/3197517.3201313

ACM Reference Format:
Zhengqin Li, Zexiang Xu, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. 2018. Learning to Reconstruct Shape and Spatially-Varying Reflectance from a Single Image. *ACM Trans. Graph.* 37, 6, Article 269 (November 2018), 11 pages. https://doi.org/10.1145/3197517.3201313

1 INTRODUCTION

Estimating the shape and reflectance properties of an object using a single image acquired "in-the-wild" is a long-standing challenge in computer vision and graphics, with applications ranging from 3D design to image editing to augmented reality. But the inherent ambiguity of the problem, whereby different combinations of shape, material and illumination might result in similar appearances, poses a significant hurdle. Consequently, early approaches have attempted to solve restricted sub-problems by imposing domain-specific priors on shape and/or reflectance [Barron and Malik 2015; Blanz and Vetter 1999; Oxholm and Nishino 2016]. Even with recent advances through deep learning based data-driven priors for inverse rendering problems, disentangling the complex factors of variation represented by arbitrary shape and spatially-varying bidirectional reflectance distribution function (SVBRDF) has, as yet, remained unsolved.

In this work, we take a step towards that goal by proposing a novel convolutional neural network (CNN) framework to estimate shape — represented as depth and surface normals — and SVBRDF — represented as diffuse albedo and specular roughness — from a single mobile phone image captured under largely uncontrolled conditions. This represents a significant advance over recent works that either consider SVBRDF estimation from near-planar samples [Aittala et al. 2016; Deschaintre et al. 2018; Li et al. 2017a, 2018], or estimate shape for Lambertian or homogeneous materials [Barron and Malik 2015;

Georgoulis et al. 2017; Liu et al. 2017]. The steep challenge of this goal requires a holistic approach that combines prudent image acquisition, a large-scale training dataset, and novel physically-motivated networks that can efficiently handle this increased complexity.

Several recent works have demonstrated that a collocated source-sensor setup leads to advantages for material estimation, since higher frequencies for specular components are easily observed and distractors such as shadows are eliminated [Aittala et al. 2016, 2015; Hui et al. 2017]. We use a mobile phone for imaging and mimic this setup by using the flash as illumination. Note that our images are captured under uncontrolled environment illumination, and not a dark room. Our only assumption is that the flash illumination is dominant, which is true for most scenarios.

Previous inverse rendering methods have utilized 3D shape repositories with homogeneous materials [Liu et al. 2017; Rematas et al. 2016; Shi et al. 2017] or large-scale SVBRDFs with near-planar geometries [Deschaintre et al. 2018; Li et al. 2018]. While we utilize the SVBRDF dataset of [Li et al. 2018], meaningfully applying them to 3D models in a shape dataset is non-trivial. Moreover, category-specific biases in repositories such as ShapeNet [Chang et al. 2015] might mitigate the generalization ability of our learned model. To overcome these limitations, we procedurally generate random shapes by combining basic shape primitives on which the complex SVBRDFs from our dataset are mapped. We generate a large-scale dataset of 180, 000 images with global illumination that reflects the distribution of flash-illuminated images under an environment map.

Besides more descriptive datasets, disambiguating shape and spatially-varying material requires novel network architectures that can reason about appearance at multiple scales, for example, to understand both local shading and non-local shadowing and lighting variations, especially in the case of unknown, complex geometry. We demonstrate that this can be achieved through a cascade design; each stage of the cascade predicts shape and SVBRDF parameters, but these predictions and the error between images rendered with these estimates and the input image are passed as inputs to subsequent stages. This allows the network to imbibe this global feedback on the rendering error, while performing iterative refinement through the stages. In experiments, we demonstrate through quantitative analysis and qualitative visualizations that the cascade structure is crucial for accurate shape and SVBRDF estimation.

The forward rendering model is well-understood in computer graphics, and can be used to aid the inverse problem by using a fixed, in-network rendering layer to render the predicted shape and material parameters and impose a "reconstruction" loss during training [Deschaintre et al. 2018; Innamorati et al. 2017; Li et al. 2018; Liu et al. 2017; Shu et al. 2017; Tewari et al. 2018]. Tractable training requires efficient rendering layers; thus, most previous works only consider appearance under direct illumination. This is insufficient, especially when dealing with arbitrary shapes. An important technical innovation of our network is a *global illumination* (GI) rendering layer that also accounts for interreflections.[1] While it is challenging to directly predict the entire indirect component of an input image, we posit that predicting the bounces of global illumination using a CNN is easier and maintains differentiability. Thus, our GI rendering is implemented as a physically-motivated cascade, where each stage predicts one subsequent bounce of global illumination. As a result, besides SVBRDF and shape, the individual bounces of global illumination are auxiliary outputs of our framework. A GI rendering layer also allows us to isolate the reconstruction error better, thereby providing more useful feedback to the cascade structure.

Contributions. In summary, we make the following contributions:

- The first approach to simultaneously recover unknown shape and SVBRDF using a single mobile phone image.
- A new large-scale dataset of images rendered with complex shapes and spatially-varying BRDF.
- A novel cascaded network architecture that allows for global reasoning and iterative refinement.
- A novel, physically-motivated global illumination rendering layer that provides more accurate reconstructions.

2 RELATED WORK

Inverse rendering — the problem of reconstructing shape, reflectance, and lighting from a set of images — is an extensively studied problem in computer vision and graphics. Traditional approaches to this problem often rely on carefully designed acquisition systems to capture multiple images under highly calibrated conditions [Debevec et al. 2000]. Significant research has also been done on the subproblems of the inverse rendering problem: e.g., photometric stereo methods that reconstruct shape assuming known reflectance and lighting [Woodham 1980], and BRDF acquisition methods that reconstruct material reflectance assuming known shape and lighting [Marschner et al. 1999; Matusik et al. 2003]. While recent works have attempted to relax these assumptions and enable inverse rendering in the "wild", to the best of our knowledge, this paper is the first to estimate both *complex shape and spatially-varying non-Lambertian reflectance from a single image captured under largely uncontrolled settings.* In this section, we focus on work that addresses shape and material estimation from sparse images.

Shape and material estimation. Shape from shading methods reconstruct shape from single images captured under calibrated illumination, though they usually assume Lambertian reflectance [Johnson and Adelson 2011]. This has been extended to arbitrary shape and reflectance under known natural illumination [Oxholm and Nishino 2016]. Shape and reflectance can also be estimated from multiple images by using differential motion cues [Chandraker 2014], light field inputs [Li et al. 2017b; Wang et al. 2017], or BRDF dictionaries [Goldman et al. 2010; Hui and Sankaranarayanan 2017]. Recent works mitigate the challenge of shape recovery by using depth maps from a Kinect sensor as input for BRDF estimation [Knecht et al. 2012; Wu and Zhou 2015]. Other methods assume near-planar samples and use physics-based optimization to acquire spatially-varying BRDFs from sparse images captured under collocated illumination [Aittala et al. 2015; Hui et al. 2017; Riviere et al. 2016]. Yu et al. [1999] assume known geometry to recover scene reflectance by modeling global illumination. Barron and Malik [2015] recover shape and spatially-varying diffuse reflectance from a single image under unknown illumination by combining an inverse rendering formulation with

[1] While it is possible to also consider shadows, global illumination is mainly manifested as interreflections in our inputs due to the collocated setup.

hand-crafted priors on shape, reflectance and lighting. In contrast to these works, our deep learning approach recovers high-quality shape and spatially-varying reflectance from a single RGB image by combining a rendering layer with purely data-driven priors.

Deep learning for inverse rendering. Recently, deep learning-based approaches have demonstrated promising results for several inverse rendering subproblems including estimating scene geometry [Bansal et al. 2016; Eigen and Fergus 2015], material classes [Bell et al. 2015], illumination [Gardner et al. 2017; Georgoulis et al. 2017; Hold-Geoffroy et al. 2017], and reflectance maps [Rematas et al. 2016]. In contrast, our work tackles the joint problem of estimating shape and spatially-varying reflectance from just a single image.

In the context of reflectance capture, Aittala et al. [2016] propose a neural style transfer approach to acquire stochastic SVBRDFs from images of near-planar samples under flash illumination. Similarly, Li et al. [2017a] acquire SVBRDFs from near-planar samples imaged under environment lighting, using a self-augmentation method to overcome the limitation of learning from a small dataset. Liu et al. [2017] propose a CNN-based method, that incorporates an in-network rendering layer, to reconstruct a homegenous BRDF and shape (from one of four possible categories) from a single image under unknown environment illumination. [Innamorati et al. 2017] use deep networks to decompose images into intrinsic components like diffuse albedo, irradiance, specular and ambient occlusion, which are recombined to specify a render loss. We use a similar render loss, though our decomposition is physically-based. Meka et al. [2018] recover homogeneous BRDF parameters of an arbitrary shape under environment lighting, and Li et al. [2018] and Deschaintre et al. [2018] leverage in-network rendering layers to reconstruct SVBRDFs from near-planar samples captured under flash illumination. Our work can be considered a generalization of all these methods — we handle a broader range of SVBRDFs and arbitrary shapes. This not only places greater demands on our network, but also necessitates the consideration of global illumination, leading to two key aspects of our architecture. First, we progressively refine shape and SVBRDF estimates through a novel cascade design that achieves large enough receptive fields while being easily trainable. Second, while previous in-network rendering layers [Deschaintre et al. 2018; Li et al. 2018; Liu et al. 2017] only consider direct illumination, our global rendering layer accounts for indirect illumination too. This not only matches our inputs better, but is also the more physically accurate choice for real scenes with complex shapes. Further, the rendering error provided as input to our cascade stages improves estimation results, which is also possible only with a rendering layer that computes global illumination. Together, these components leads to state-of-the-art results on a significantly broader range of inputs.

Rendering layers in deep networks. Differentiable rendering layers have been used to aid in the task of learning inverse rendering for problems like face reconstruction [Sengupta et al. 2018; Shu et al. 2017; Tewari et al. 2018] and material capture [Deschaintre et al. 2018; Li et al. 2018; Liu et al. 2017]. However, these methods make simplifying assumptions — usually Lambertian materials under distant direct lighting — to make these layers tractable. We also use rendering to introduce information from varied lighting conditions, but in contrast to the above works, our rendering accounts for global illumination. Since analytical rendering of global illumination is challenging, we rely on network modules to predict bounces of global illumination. The idea of using a network to predict global illumination has also been adopted by [Nalbach et al. 2017], but no prior method has done this for inverse problems. Further, we use a physically meaningful network structure that divides global illumination into several bounces instead of directly predicting indirect lighting, which may lead to better and more interpretable results. A deep network is also used by [Marco et al. 2017] to compensate for global illumination in time-of-flight measurements, but they use a black box network for depth prediction while we model global illumination explicitly. There is machinery to compensate for bounces in optimization-based methods [Godard et al. 2015], but they do not render in real-time and there is no obvious way to back-propagate gradients, making them unsuitable for our framework. We train a global illumination CNN to predict multiple bounces using data generated using a novel simulation-based strategy that renders random shapes with a large-scale SVBRDF dataset. The use of random shapes is important, since we aim to recover arbitrary geometry, unlike previous methods that might incorporate semantic category-level priors [Chang et al. 2015; Georgoulis et al. 2017; Liu et al. 2017; Meka et al. 2018; Rematas et al. 2016]. Besides higher accuracy in SVBRDF estimation, a collateral benefit of our novel rendering layer is that it can predict individual bounces of global illumination, in the same forward pass. These can be subsequently used for scene analysis tasks [Nayar et al. 2006; O'Toole and Kutulakos 2010].

Cascade networks. For prediction tasks that demand sub-pixel accuracy, prior works have considered cascade networks. For instance, convolutional pose machines [Wei et al. 2016] are devised to obtain large receptive fields for localizing human body joints, while other architectures such as deep pose [Toshev and Szegedy 2014] and stacked hourglass networks [Newell et al. 2016] also use cascades for multiscale refinement. Improved optical flow estimates are obtained by FlowNet 2.0 [Ilg et al. 2017] using cascaded FlowNet modules that accept stage-wise brightness error as input. Similar to the above, we show that the cascade structure is effective for SVBRDF estimation. Uniquely, we demonstrate that our cascade is sufficient to recover high-quality shape and SVBRDF, while our global illumination prediction that enables rendering error as input to the cascade stages also yields advantages for SVBRDF estimation.

3 METHOD

The input to our method is a single image of an object (with a mask) captured under (dominant) flash and environment illumination. Reconstructing spatially-varying BRDF (SVBRDF) and shape, in such uncontrolled settings, is an extremely ill-posed problem. Inspired by the recent success of deep learning methods in computer vision and computer graphics, we handle this problem by training a CNN specifically designed with intuition from physics-based methods. In this section, we will describe each component of our network. The overall framework is shown in Figure 2.

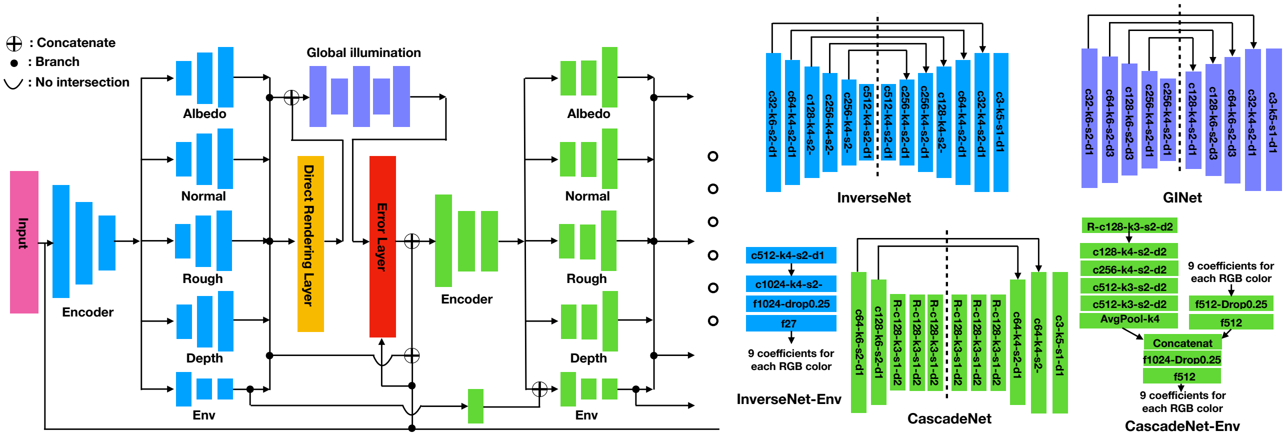


Fig. 2. **Right:** Overall structure of our framework. Different colors specify different functions of the network (blue for initial estimation, green for refinement and purple for global illumination prediction). We use a cascade of encoder-decoder networks for global reasoning and iterative refinement. Different cascade levels do not share parameters since the input statistics at each stage and the refinements needed are different. Each cascade stage receives error feedback through the rendered output of the previous stage. Since we handle arbitrary shapes, our rendering layer models individual bounces of global illumination. **Left:** Details of hyperparameters in our physically-motivated network design. Here R represents a residual block [He et al. 2016]. $\mathrm{c}X_1-\mathrm{k}X_2-\mathrm{s}X_3-\mathrm{d}X_4$ represents a conv/deconv layer of output channel X_1, kernel size X_2, stride X_3 and dilation X_4. Our encoder has receptive fields large enough to model global light illumination, skip links are added since we aim to recover fine details and large kernels are used for global illumination prediction.

3.1 Basic Architecture

Our basic network architecture consists of a single encoder and four decoders for different shape and SVBRDF parameters: diffuse albedo (A), specular roughness (R), surface normal (N), and depth (D).[2] For simplicity, we start by considering the input to be an image, I^p, of an object illuminated by a dominant point light source collocated with the camera (we consider additional environment illumination in Section 3.3). We manually create a mask, M, that we stack with the image to form a four channel input for the encoder. A light source collocated with the camera has the advantages of removing cast shadows, simplifying the lighting conditions and easing observation of high frequency specularities, which are crucial for solving the inverse rendering problem. In our experiments, such input data is easily acquired using a mobile phone with the flash light enabled. Unlike [Li et al. 2017a], which has different encoders and decoders for various BRDF parameters, our four decoders share features extracted from the same encoder. The intuition behind this choice is that different shape and SVBRDF parameters are closely correlated, thus, sharing features can greatly reduce the size of the network and alleviate over-fitting. This architecture has been proven to be successful in [Li et al. 2018] for material capture using near-planar samples. Let **InverseNet**$(\cdot)$ be the basic network architecture consisting of the encoder-decoder block (shown in blue in Figure 2). Then the initial predicted shape and SVBRDF estimates (differentiated from the true parameters by ˜) are given by:

$$\tilde{A}, \tilde{N}, \tilde{R}, \tilde{D} = \mathbf{InverseNet}(I^p, M). \tag{1}$$

3.2 Global Illumination Rendering Layer

Prior works on material capture or photometric stereo usually assume that the influence of inter-reflections can be neglected, or consider near-planar samples where its effects are not strong. However, that may not be the case for our setup, since we consider complex shape with potentially glossy reflectance. Failing to model global illumination for our problem can result in color bleeding and flattened normal artifacts. We initially considered in-network global illumination rendering during training, but found it time-consuming and not feasible for a large dataset. Instead we propose using CNNs to approximate global illumination. CNNs can capture the highly non-linear operations that global illumination manifests. In addition, they have the advantage of being differentiable and fast to evaluate.

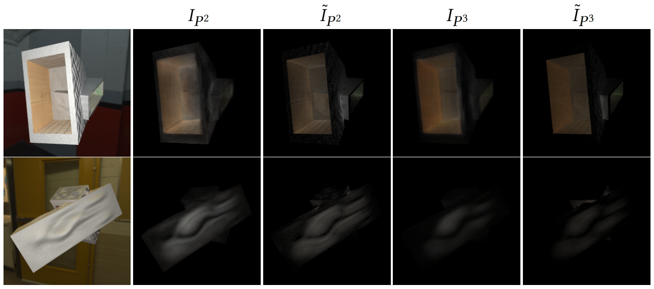
In particular, we use a series of CNNs, each of which predict individual bounces of the rendered image. Let $\mathbf{GINet}_n$ be the n-bounce CNN. This network is trained to takes the $(n-1)$-bounce image under point light illumination, I^p_{n-1}, and the shape and SVBRDF parameters, and render the n-bounce image, $\tilde{I}^p_n$, as:

$$\tilde{I}^p_n = \mathbf{GINet}_n(I^p_{n-1}, M, A, N, R, D) \tag{2}$$

We use an analytical rendering layer to compute the direct illumination, i.e., first bounce image, $\tilde{I}^p_1$, given the predicted shape and SVBRDF parameters. Then we use two CNNs, $\mathbf{GINet}_2(\cdot)$ and $\mathbf{GINet}_3(\cdot)$, to predict the second and third bounces, $\tilde{I}^p_2$ and $\tilde{I}^p_3$ respectively. The output, $\tilde{I}^p_g$, of our full global illumination rendering layer (shown in purple in Figure 2) sums all the bounce images as:

$$\begin{aligned} \tilde{I}^p_2 &= \mathbf{GINet}_2(\tilde{I}^p_1, M, \tilde{A}, \tilde{N}, \tilde{R}, \tilde{D}), \\ \tilde{I}^p_3 &= \mathbf{GINet}_3(\tilde{I}^p_2, M, \tilde{A}, \tilde{N}, \tilde{R}, \tilde{D}), \\ \tilde{I}^p_g &= \tilde{I}^p_1 + \tilde{I}^p_2 + \tilde{I}^p_3. \end{aligned} \tag{3}$$

[2] A specular albedo may be considered too, but we found it sufficient to consider just roughness to mimic most real-world appearances.



	1^{st} bounce	2^{nd} bounce	3^{rd} bounce	others
Energy ratio	95.83%	3.08%	0.89%	0.20 %

Fig. 3. Global illumination prediction results. From left to the right are input images, the predicted second bounce images, the ground truth second bounce images, the predicted third bounce images and the ground truth third bounce images. Even for complex shapes with glossy material, the predictions of our network are close to the ground truth. On the bottom, we show the ratio between the average energy of separate bounces and the images illuminated by a point light source across the test dataset.

As illustrated in Figure 3, most of the image intensity is contained within three bounces, and so we only predict these, ignoring subsequent bounces. Also in Figure 3, we show second and the third bounce images predicted by our network. We observe that even for objects with very concave shape and highly glossy material, we can still generate rendering outputs that closely match the ground truth.

Note that our CNN-based global illumination network only approximates true global illumination. It operates in image space and does not explicitly model interreflections from surface points that are not visible to the camera. However, our training data does include the interreflections from invisible surfaces and our collocated setup up causes interreflections from visible regions to dominate. In practice, we have found the network to be sufficiently accurate for inverse rendering. Compared with the traditional radiosity method [Cohen and Wallace 1993], our network-based global illumination prediction has the advantage of being fast, differentiable and able to approximate the reflection from invisible surfaces. However, it is an approximation, since we do not have precise geometry, form factors or material (albedo) properties, as in conventional radiosity algorithms.

3.3 Environment Map Prediction

Although we use a dominant flash light, our images are also illuminated by unknown environment illumination. This environment illumination can significantly affect the appearance of globally illuminated complex shapes. This requires us to estimate the environment illumination and account for it in our rendering networks. To do so, we approximate environment lighting with low-frequency spherical harmonics (SH), and add another branch to our encoder-decoder structure to predict the first nine SH coefficents for each color channel. We observe that the background image provides important context information for the network to determine environment lighting. So, unlike the point light source case, we add the image with background as the third image to the input. Let E be environment lighting, I^{pe} be the image of the object under both point and environmental lighting and $M \odot I^{pe}$ be its masked version. With some abuse of notation, now our shape and SVBRDF parameters are computed using

$$\tilde{A}, \tilde{N}, \tilde{R}, \tilde{D}, \tilde{E} = \mathbf{InverseNet}(I^{pe}, M \odot I^{pe}, M). \qquad (4)$$

Since now the input image is captured under environment illumination and the flash light source, we modify our rendering layer to account for this. We follow the method of [Ramamoorthi and Hanrahan 2001] to render an image of the object, $\tilde{I}_e$, using the estimated spherical harmonics illumination. This only considers the Lambertian shading and ignores high-frequency specular effects. In practice, this is sufficient because most high-frequency effects are observed under flash illumination, and our experiments show that this simple approximation suffices for achieving accurate BRDF reconstruction. Now the output of the global illumination rendering layer (in place of Equation 3) is given by:

$$\tilde{I}^{pe}_g = \tilde{I}^p_1 + \tilde{I}^p_2 + \tilde{I}^p_3 + \tilde{I}^e. \qquad (5)$$

3.4 Cascade Structure

While a single encoder-decoder leads to good results for SVBRDF estimation with near-planar samples [Li et al. 2018], it does not suffice when considering arbitrary shapes. This can be attributed to the increased complexity of the problem and a need for more global reasoning. We propose a cascade structure that achieves these aims by using iterative refinement and feedback to allow the network to reason about differences between the image rendered with the predicted parameters and the input image.

Let $\mathbf{CascadeNet}_n$ be stage n of the cascade network. Each stage has the same single encoder-four decoders architecture as **InverseNet**. Let the shape, reflectance and lighting parameters of cascade stage n be $\tilde{A}_n$, $\tilde{N}_n$, $\tilde{R}_n$, $\tilde{D}_n$ and $\tilde{E}_n$, and the result of rendering these parameters (using the global illumination rendering network) be $\tilde{I}^{pe}_{g,n}$. Each cascade stage refines the predictions of the *previous* stage as:

$$\begin{aligned} Err_{n-1} &= M \odot I^{pe} - \tilde{I}^{pe}_{g,n-1} & (6) \\ \tilde{A}_n, \tilde{N}_n, \tilde{R}_n, \tilde{D}_n, \tilde{E}_n &= \mathbf{CascadeNet}_n(I^{pe}, M \odot I^{pe}, M, \\ & \qquad \tilde{A}_{n-1}, \tilde{N}_{n-1}, \tilde{R}_{n-1}, \tilde{D}_{n-1}, Err_{n-1}) & (7) \end{aligned}$$

The inputs to each cascade stage are the input image, the shape, SVBRDF, and lighting predictions from the previous stage, and the rendering error associated with these previous predictions (with respect to the input image). This allows each cascade stage to refine the predictions by reasoning about the rendering error from the previous stage. Note that this is possible only because of our network design that models global illumination and environment lighting.

3.5 Training Details

Training Data: To the best of our knowledge, there is no existing dataset of objects with arbitrary shape rendered with complex SVBDRF. Complex SVBRDF datasets used in previous work [Deschaintre et al. 2018; Li et al. 2018] assume near-planar surfaces, and rich shape datasets like ShapeNet [Chang et al. 2015] have simple homogeneous BRDFs. Thus, we generate our own synthetic dataset by procedurally adding shapes to build a complex scene. Similar to [Xu et al. 2018], we first generate primitive shapes (cube, ellipsoid, cylinder, box and L-shape) and then add a randomly generated height map to make them more complex and diverse. We build scenes by randomly sampling 1 to 5 shapes and combining them. We create 3600 scenes, using 3000 for training and 600 for testing.

We use SVBDRFs from the Adobe Stock material dataset[3], which contains 694 complex SVBRDFs spanning a large range of material types. Each SVBRDF is comprised of 4K texture maps for diffuse albedo, specular roughness, and surface normals, and uses the physically motivated microfacet BRDF model described in [Karis and Games 2013]. Please refer to the supplementary material for details.

We remove the 6 transparent materials and use the remaining 688 materials. We classify the materials into 8 categories according to their reflectance properties and proportionally sample 588 materials for training and 100 for testing. For environment maps, we use the Laval Indoor HDR dataset [Gardner et al. 2017] containing 2144 environmental maps of indoor scenes, of which we use 1500 to render the training dataset and 644 for the test dataset.

We use Optix for GPU-accelerated rendering, based on path tracing with multiple importance sampling. We render with 400 samples per-pixel for point light source illumination and 625 samples per-pixel when the environment map is also included. The average rendering time is less than 2 seconds. For each scene, we sample 12 viewing directions, 5 groups of different SVBDRFs and one environment map. When rendered with both point and environment lighting, we scale the environment map by 0.5, to keep the average ratio between image intensities rendered with only environment map and with point light to be 0.09285. This matches the statistics of images captured using mobile phones in real indoor environments.

Network Design: Our design makes several choices to reflect the physical structure of the problem. We use the U-net architecture [Ronneberger et al. 2015] for **InverseNet**. To model the global fall-off of the point light source, it is necessary to have large receptive fields. Thus, each encoder has 6 convolutional layers with stride 2, so that each pixel of the output can be influenced by the whole image. For the SVBDRF parameters, we use transposed convolutions for decoding and add skip links to recover greater details. For environment map estimation, we pass the highest level of feature extracted from the encoder through two fully connected layers to regress the 9 spherical harmonics coefficients. Each **CascadeNet** stage uses 6 residual blocks — 3 blocks for the encoder and 3 separate blocks for each decoder. We use dilated convolutions with a factor of 2 in the residual block to increase the receptive field. We feed environment lighting predictions into the next cascade stage by passing the nine SH coefficients through a fully connected layer and concatenate them with the feature extracted from the encoder. We also use the U-net structure with skip-links for **GINet**. To predict global illumination, the network must capture long range dependencies. Thus, we use a convolutional layer with large kernel of size 6, combined with dilation by a factor of 2. The network architecture of each component is shown on the right side of Figure 2.

Loss function: We have the same loss function for both **InverseNet** and each **CascadeNet** stage. For diffuse albedo, normal, roughness and environment illumination SH coefficients, we use the L2 loss for supervision. Since the range of depths is larger than that of other BRDF parameters, we use an inverse transformation to project the depth map into a fixed range. Let $\tilde{d}_i$ be the initial output of depth prediction network of pixel i; the real depth d_i is given by

$$d_i = \frac{1}{\sigma \cdot (\tilde{d}_i + 1) + \epsilon}. \tag{8}$$

We set $\sigma = 0.4$ and $\epsilon = 0.25$, and use L2 loss to supervise d_i. Finally, we add a reconstruction loss based on the L2 distance between the image rendered with predicted and ground truth parameters. Let $\mathcal{L}_a$, $\mathcal{L}_n$, $\mathcal{L}_r$, $\mathcal{L}_d$, $\mathcal{L}_{env}$ and $\mathcal{L}_{rec}$ be the L2 losses for diffuse albedo, normal, roughness, depth, environment map and image reconstruction, respectively. The loss function of our network is:

$$\mathcal{L} = \lambda_a \mathcal{L}_a + \lambda_n \mathcal{L}_n + \lambda_r \mathcal{L}_r + \lambda_d \mathcal{L}_d + \lambda_{env} \mathcal{L}_{env} + \lambda_{rec} \mathcal{L}_{rec}, \tag{9}$$

where $\lambda_a = \lambda_n = \lambda_{rec} = 1$, $\lambda_r = \lambda_d = 0.5$ and $\lambda_{env} = 0.1$ are parameters chosen empirically.

Training Strategies: Training multiple cascade structures is difficult since the enhanced network depth may lead to vanishing gradients and covariate shift, preventing convergence to a good local minimum. Further, batch sizes will need to be small when training all stages together, which can cause instability. Thus, instead of training the whole network end-to-end, we sequentially train each stage of the cascade. This allow us to use a relatively large batch size of 16. We use Adam optimizer, with a learning rate of 10^{-4} for the encoder and 4×10^{-4} for the decoders. We decrease the learning rate by half after every two epochs. The three stages are trained for 15, 8 and 6 epochs, respectively. We use two **CascadeNet** stages and train **InverseNet** and **CascadeNet**$_1$ with 2500 shapes and add 500 shapes to train **CascadeNet**$_2$.

GINet is trained prior to the BRDF prediction network, then held fixed and only used for the rendering layer when training the network for shape and SVBRDF estimation. We use Optix to render images with separate bounces and use them for direct supervision. We train for 15 epochs, with an initial learning rate of 2×10^{-4} and reduce it by half every two epochs.

4 EXPERIMENTS

We first demonstrate the effectiveness of each design choice in our network architecture through detailed comparisons on both synthetic and real datasets. Next, we compare with previous methods for shape and material estimation to highlight the highly accurate shape and SVBRDF recovered by our framework. Please refer to supplementary video for more visualizations of the results (including under novel lighting and viewpoint).

Ablation study on synthetic data. We first justify the necessity of rendering a novel large-scale dataset with global illumination for shape and SVBRDF estimation. We train **InverseNet** on images rendered with direct illumination and test on images with global illumination. Column $\texttt{Im}^{\texttt{p}}_{\texttt{d}}\texttt{-C0}$ (trained on images with direct point illumination with no cascade) in Table 1 reports the obtained errors, which are clearly larger than those in column $\texttt{Im}^{\texttt{p}}_{\texttt{g}}\texttt{-C0}$ for the same network trained on images with point lighting and global illumination. Thus, global illumination has a significant impact on depth and

[3] https://stock.adobe.com/3d-assets

Table 1. Quantitative comparison on images rendered only with point light. $\mathtt{Im}_d^{pe}$ refers to input images rendered with direct lighting only, while $\mathtt{Im}_g^{pe}$ means the input images are rendered with global illumination.

	$\mathtt{Im}_d^{p}-\mathtt{C}_0$	$\mathtt{Im}_g^{p}-\mathtt{C}_0$
Albedo(10^{-2})	5.911	5.703
Normal(10^{-2})	4.814	4.475
Roughness(10^{-1})	1.974	1.966
Depth(10^{-1})	1.842	1.772

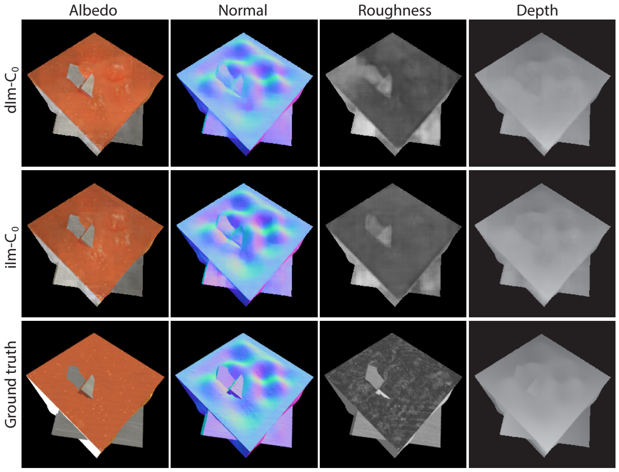


Fig. 4. Comparison of SVBRDF and depth outputs of two networks, trained on directly illuminated (top) and globally illuminated images (middle), when evaluated on an input with global illumination. Not considering indirect lighting during training leads to flatter normals and brighter albedo.

SVBRDF estimation. The qualitative comparison in Figure 4 shows that the network trained with direct lighting only predicts brighter diffuse albedo and flattened normals, when evaluated on images with indirect lighting. This also matches intuition on the behavior of inter-reflections [Chandraker et al. 2005; Nayar et al. 1991].

Next we demonstrate that context information is important for the network to reconstruct shape and BRDF under environment lighting. We train two variants of our basic network, one with masked image input, $\mathtt{Im}_g^{pe}-\mathtt{C}_0$, and the other with both masked and original image as input, $\mathtt{Im}_g^{pe}-\mathtt{bg}-\mathtt{C}_0$. Quantitative comparisons in the first two columns of Table 2 show that predictions for all BRDF parameters improve when background is included.

To test the effectiveness of cascade structure, we first add one layer of cascade to our basic network. We try two variants of cascade network. For the black-box cascade ($\mathtt{C}_1$), we stack the input image and the predicted BRDF parameters and send them to the next stage of the cascade. For the cascade network with error feedback ($\mathtt{C}_1\mathtt{Er}$), we also send an error map as input by comparing the output of our global illumination rendering layer with the input. The quantitative numbers (third and fourth column of Table 2) suggest that having the error feedback improves BRDF reconstruction. We then add another cascade stage with error feedback, which yields even more accurate BRDF estimation ($\mathtt{C}_2\mathtt{Er}$) that we deem the final output.

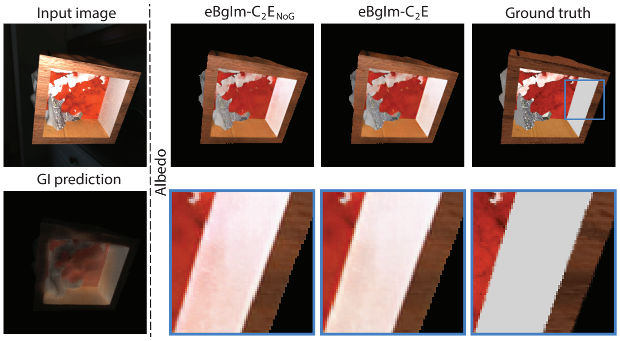


Fig. 5. For an input image with strong indirect lighting (top left), a network trained without global illumination for the rendering layer (second column) retains more color bleeding artifacts in the estimated diffuse albedo, than one trained with global illumination (third column). The bottom left figure shows the net global illumination estimated by the final network.

Figure 6 shows the visual quality of BRDF estimation from different stages of the cascade network. We observe that for both synthetic and real data, the cascade reduces noise and artifacts. The final rendered result using the BRDF parameters predicted by the second level of the cascade is very similar to the input image, as shown in Figure 6 using both the environment map estimated by the network and a novel environment map.

Next, we analyze the effect of the global illumination rendering network. We train two new variants of our global illumination rendering layer for the second cascade stage. For $\mathtt{Im}_g^{pe}-\mathtt{bg}-\mathtt{C}_2\mathtt{Er}_{\mathtt{NoG}}$, the rendering layer does not consider global illumination so that the error feedback is computed by subtracting the sum of $\tilde{I}_1^p$ and $\tilde{I}^e$ from the input $M \odot I^{pe}$, i.e., $\tilde{I}^{pe} = \tilde{I}_1^p + \tilde{I}^e$. Similarly, for $\mathtt{Im}_g^{pe}-\mathtt{bg}-\mathtt{C}_2\mathtt{Er}_{\mathtt{NoE}}$, we remove the environmental map component of the global illumination rendering layer. The error feedback for the cascade network is now computed using $\tilde{I}^{pe} = \tilde{I}_1^p + \tilde{I}_2^p + \tilde{I}_3^p$. Table 2 shows that our full version of rendering layer performs the best. The differences are measurable but subtle, since the remaining impact of environment lighting and global illumination for the second stage is small. To better understand the behavior, we show a qualitative example with global illumination in Figure 5. We observe that the global illumination rendering layer alleviates color bleeding artifacts.

Generalization to real data. We demonstrate our method on several real objects in Figures 7 and 8. All images are captured in indoor scenes using an iPhone 10 with the flash enabled. We use the Adobe Lightroom app to capture linear images and manually create the segmentation mask. For all the examples, our rendered results closely match the input. Figures 7 and 8 also show our predicted BRDF parameters can be used to render realistic images under new environment lighting and camera pose. This demonstrates that our estimates of the surface normal and spatially varying roughness are of high enough quality to render realistic specular effects of real objects under novel illumination and viewing directions.

Comparisons with previous methods. Since we are not aware of prior works that can use a single image for spatially varying BRDF

Table 2. Quantitative comparisons L2 errors illustrating the influence of various network choices. $\mathtt{Im}_g^{pe}$ means the input images are illuminated by both point light source and environmental lighting (superscript pe) and rendered with global illumination (subscript g). $-$bg means the images without masking the background are added as an input. $\mathtt{C}_n$ shows the level of cascade refinement, where $\mathtt{C}_\emptyset$ means we use our basic InverseNet without any refinement. $\mathtt{Er}$ behind $\mathtt{C}_n$ means we also send the error maps by comparing the images rendered with the estimated BRDFs and the inputs to the cascade refinement networks. The subscript $\mathtt{NoE}$ and $\mathtt{NoG}$ in the last two columns means that when computing the error maps, we do not consider the influence of environmental lighting and global illumination respectively. Here, $\mathtt{Im}_g^{pe}-\mathtt{bg}-\mathtt{C}_2\mathtt{Er}$ is the error obtained with our final two-cascade architecture with global illumination and error feedback.

	$\mathtt{Im}_g^{pe}-\mathtt{C}_\emptyset$	$\mathtt{Im}_g^{pe}-\mathtt{bg}-\mathtt{C}_\emptyset$	$\mathtt{Im}_g^{pe}-\mathtt{bg}-\mathtt{C}_1$	$\mathtt{Im}_g^{pe}-\mathtt{bg}-\mathtt{C}_1\mathtt{Er}$	$\mathtt{Im}_g^{pe}-\mathtt{bg}-\mathtt{C}_2\mathtt{Er}$	$\mathtt{Im}_g^{pe}-\mathtt{bg}-\mathtt{C}_2\mathtt{Er}_{\mathtt{NoE}}$	$\mathtt{Im}_g^{pe}-\mathtt{bg}-\mathtt{C}_2\mathtt{Er}_{\mathtt{NoG}}$
Albedo(10^{-2})	6.089	5.670	5.150	5.132	**4.868**	4.900	4.880
Normal(10^{-2})	4.727	4.580	3.929	3.907	**3.822**	3.830	**3.822**
Roughness(10^{-1})	2.207	2.064	2.004	2.011	**1.943**	1.948	1.947
Depth(10^{-2})	1.945	1.871	1.631	1.624	**1.505**	1.512	1.511
Bounce 1(10^{-3})	3.526	3.291	2.190	2.046	**1.637**	1.643	1.643
Bounce 2(10^{-4})	2.88	2.76	2.47	2.47	**2.45**	2.45	2.46
Bounce 3(10^{-5})	6.6	6.4	5.9	5.9 .	**5.8**	**5.8**	**5.8**

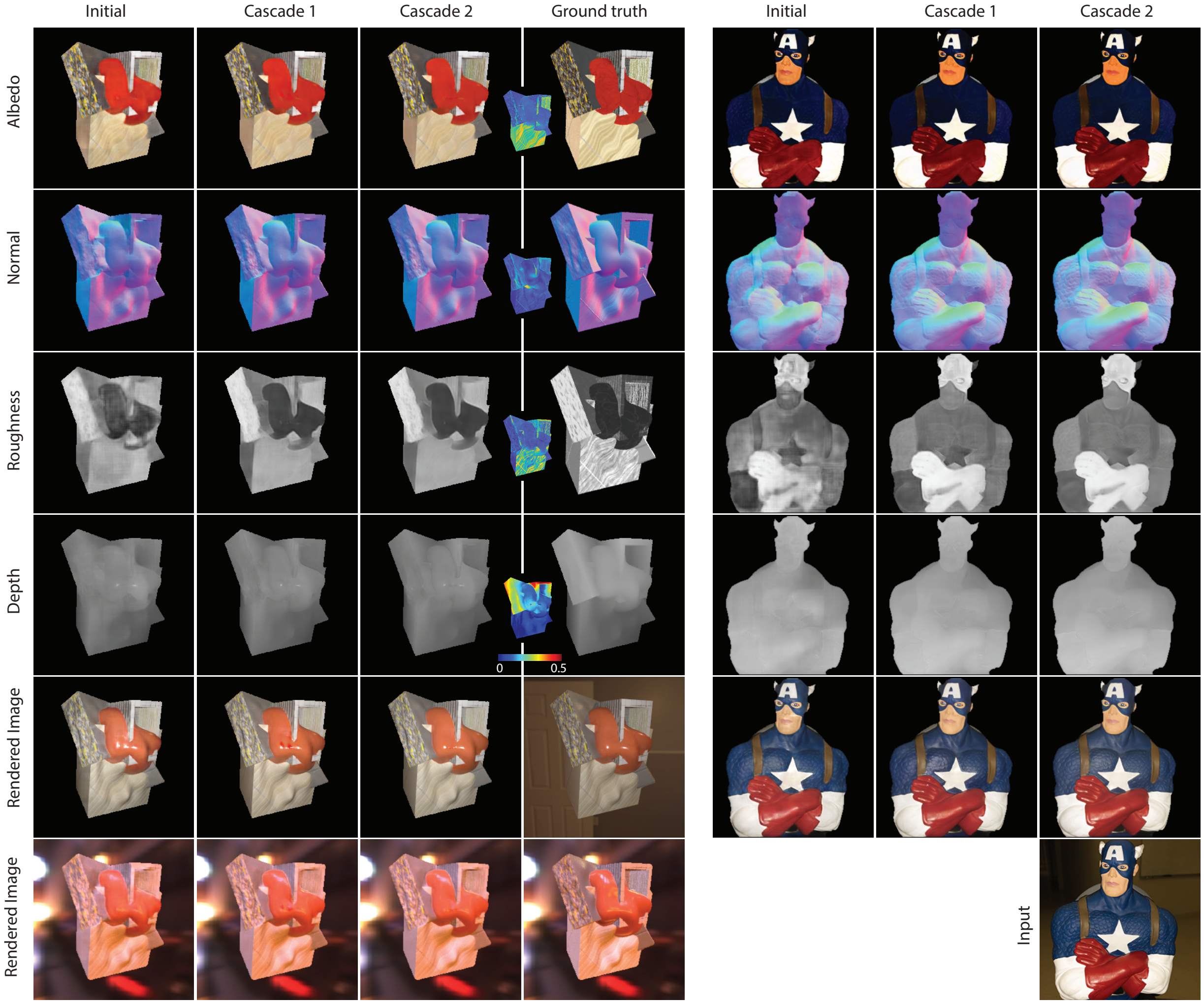


Fig. 6. Effect of our cascaded design, illustrated for synthetic (left) and real data (right). It is observed that predictions from the initial network are somewhat inaccurate, but progressively improved by the cascade stages. Images rendered after two cascade stages have less artifacts and display specular highlights closer to the ground truth, both when relit with the estimated environment map and rendered under a new environment map. We visualize the absolute error for the BRDF parameters in the third column except the depth error. The depth error is normalized so that the range of ground-truth depth is 1.

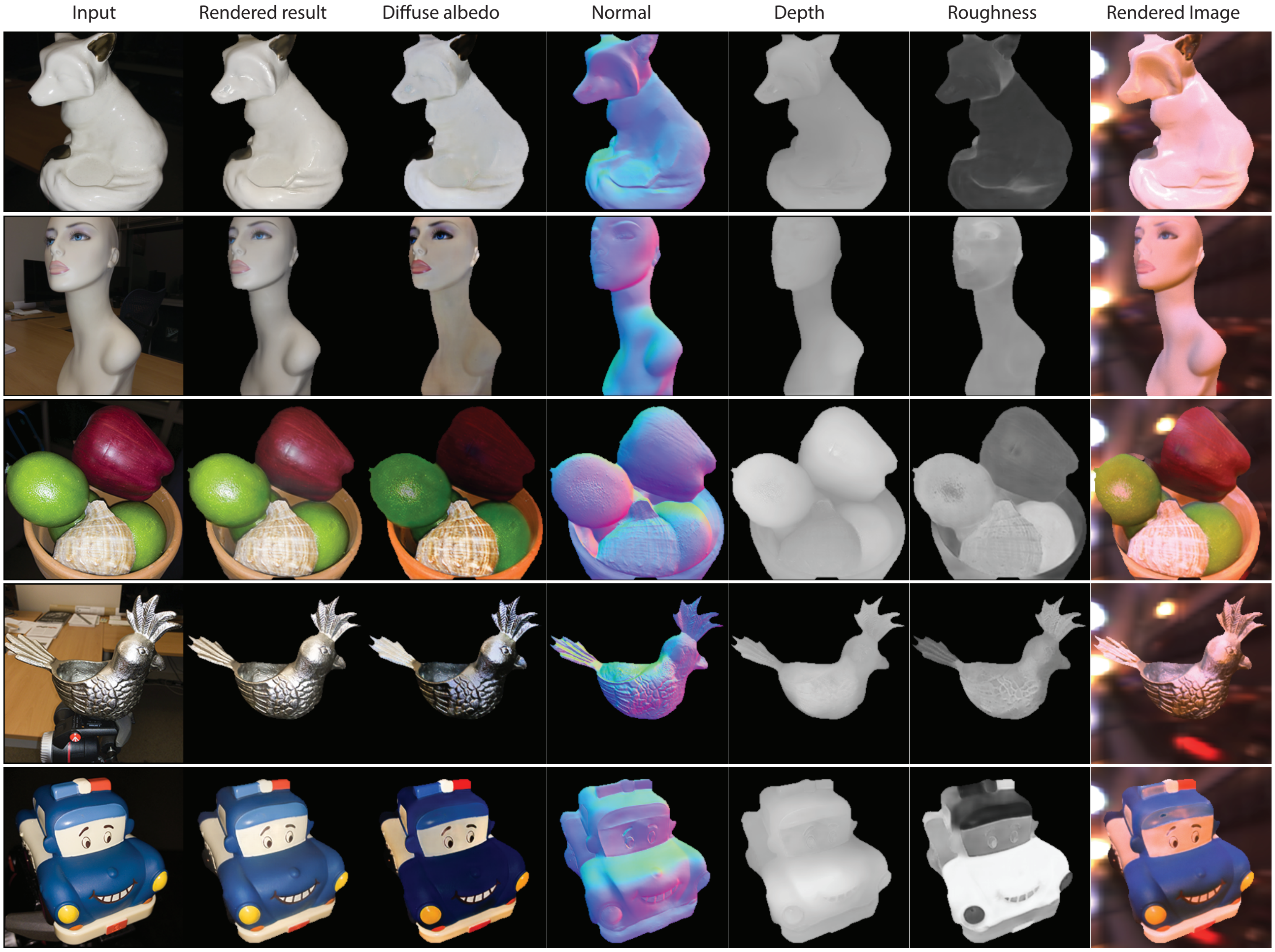


Fig. 7. Results on real objects. For each example, we show the input image, the rendered output using the estimated shape and BRDF parameters, as well as visualization under a novel illumination condition. In each case, we observe high quality recovery of shape and spatially-varying BRDF.

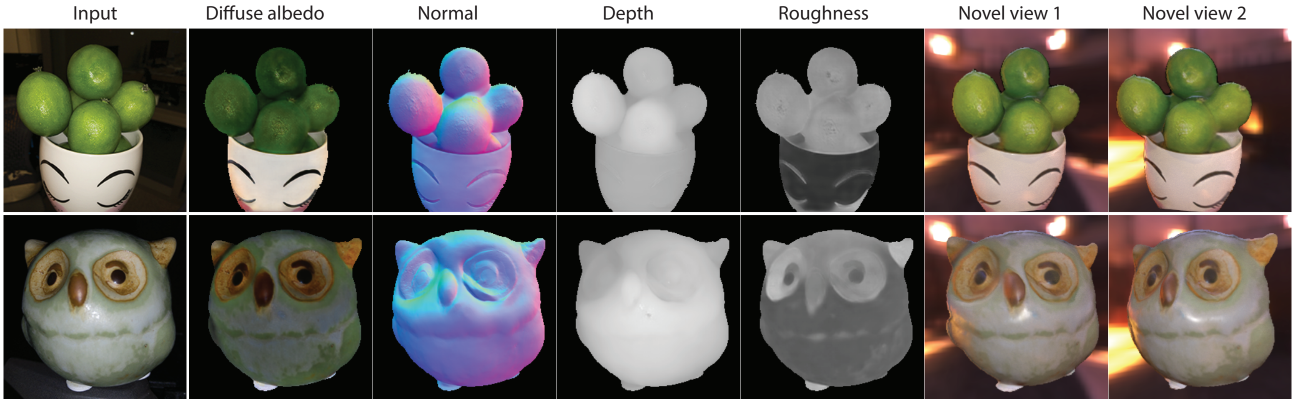


Fig. 8. Results rendered from novel views. We show the input image, the estimated shape and BRDF parameters and the rendered output under an environment map from two novel views. We observe high fidelity rendered images, as well as high quality recovery of shape and spatially-varying BRDF.

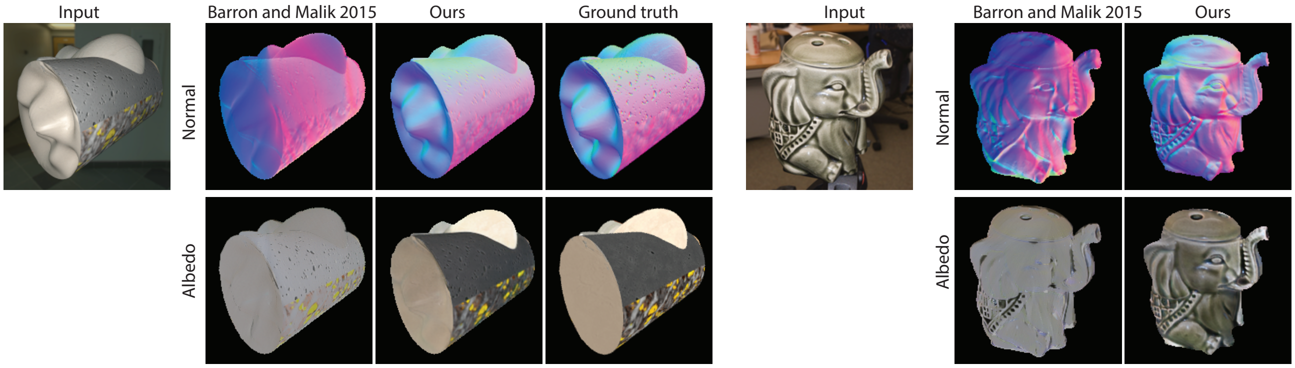


Fig. 9. Comparison with SIRFS [Barron and Malik 2015]. Our method accurately estimates the shape and diffuse color, even in regions with specularity. In contrast, because of the complex shape and materials of these objects, SIRFS, which assumes Lambertian reflectance, produces very inaccurate estimates.

and shape estimation, our comparisons are to more restricted methods for shape and material estimation, or to intrinsic image decomposition methods. We first compare with SIRFS [Barron and Malik 2015] which jointly reconstructs shape and diffuse color. Figure 9 compares the diffuse albedo and normal estimated using SIRFS with those obtained by our framework, on both real and synthetic data. In both cases, our estimates are significantly better. Notice that SIRFS tends to over-smooth both the diffuse color and the normal due to a handcrafted regularization. In contrast, our method successfully recovers high-frequency details for both diffuse albedo and surface normals, even in specular and shadowed regions.

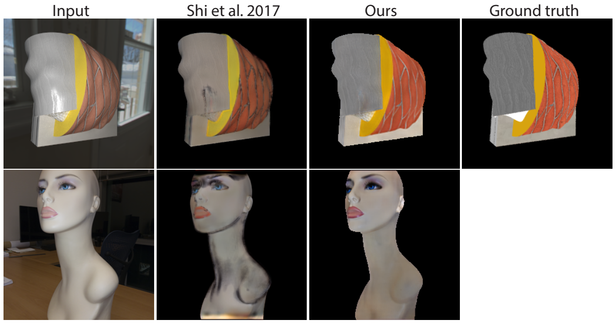


Fig. 10. Comparison with [Shi et al. 2017]. While Shi et al. train to handle non-Lambertian reflectance, the accuracy and visual quality of our diffuse albedo is significantly higher on both synthetic (top) and real data (bottom).

We also compare with the recent intrinsic image decomposition method of [Shi et al. 2017], which is trained to separate diffuse and specular components from a single image of a ShapeNet object [Chang et al. 2015], rendered under the assumption of a parametric homogeneous BRDF. We compare to their diffuse albedo prediction in Figure 10. Our method can better preserve occlusion boundaries and recover accurate diffuse color even in specular regions. Our method also yields qualitatively superior results on real data.

Limitations. A few challenges remain unaddressed. Our network does not explicitly handle improperly exposed images. For example, saturations from the flash may cause the specular highlight to be baked into the diffuse color (such as the orange in the third row of Figure 7). This problem might be solved by adding more training data and using more aggressive data augmentation. Also, since we approximate global illumination with an image-space CNN, long-range interactions might not be sufficiently modeled, which may limit the CNN's ability in correctly handling the interreflection. We find spatially varying roughness prediction to be an extremely challenging problem. The presence of specular highlights is important for roughness prediction. When there is no specular highlight, the network may rely on a long range connectivity prior to predict roughness. However, this prior may fail, which results in the same material having different roughness values (such as the owl in the second row of Figure 8). Such a prior might be explicitly enhanced to improve the performance by using a densely connected CRF [Ristovski et al. 2013] or bilateral filter [Barron and Poole 2016]. From Figure 6, we can see that the error of depth prediction is significantly larger than the normal prediction, which suggests that we may use normal predictions to refine depth predictions [Nehab et al. 2005]. We do not model global illumination from invisible surfaces, but the network does learn from training data which includes their effects and visible interreflections dominate in our collocated setup. Despite these limitations, we note our network achieves significantly better results than prior works on this challenging, ill-posed problem.

5 CONCLUSION

We demonstrate the first approach for simultaneous estimation of arbitrary shape and spatially-varying BRDF, using a single mobile phone image. We make several physically-motivated and effective choices across image acquisition, dataset creation and network architecture. We use a mobile phone flash to acquire images, which allows observing high frequency details. Our large-scale dataset of procedurally created shapes, rendered with spatially-varying BRDF under various lighting conditions, prevents entanglement of category-level shape information with material properties. Our cascaded network allows global reasoning through error feedback and multiscale iterative refinement, to obtain highly accurate outputs for both shape and material. We propose a novel rendering layer to incorporate information from various lighting conditions, which must account for global illumination to handle arbitrary shape. Inspired by the physical process of rendering bounces of global illumination, we devise a cascaded CNN module that retains speed and simplicity. Extensive experiments validate our network design through high-quality estimation of shape and SVBRDF that outperforms previous methods. In future work, we will demonstrate applications of our framework to material editing and augmented reality, as well as consider extensions to large-scale scenes such as room interiors.

ACKNOWLEDGEMENTS

We thank the reviewers for many helpful suggestions. This work was supported in part by NSF CAREER 1751365, NSF Grant 1703957, ONR grant N000141712687, Adobe, the Ronald L. Graham Chair, and the UC San Diego Center for Visual Computing.

REFERENCES

Miika Aittala, Timo Aila, and Jaakko Lehtinen. 2016. Reflectance modeling by neural texture synthesis. *ACM Trans. Graphics* 35, 4 (2016).

Miika Aittala, Tim Weyrich, Jaakko Lehtinen, et al. 2015. Two-shot SVBRDF capture for stationary materials. *ACM Trans. Graphics* 34, 4 (2015).

Aayush Bansal, Bryan Russell, and Abhinav Gupta. 2016. Marr Revisited: 2D-3D Model Alignment via Surface Normal Prediction. In *CVPR*.

Jonathan T Barron and Jitendra Malik. 2015. Shape, illumination, and reflectance from shading. *PAMI* 37, 8 (2015).

Jonathan T Barron and Ben Poole. 2016. The fast bilateral solver. In *European Conference on Computer Vision*. Springer, 617–632.

Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. 2015. Material Recognition in the Wild with the Materials in Context Database. In *CVPR*.

Volker Blanz and Thomas Vetter. 1999. A morphable model for the synthesis of 3D faces. In *Proc. SIGGRAPH*.

Manmohan Chandraker. 2014. On shape and material recovery from motion. In *ECCV*.

Manmohan Chandraker, Fredrik Kahl, and David Kriegman. 2005. Reflections on the generalized bas-relief ambiguity. In *CVPR*.

Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. 2015. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012* (2015).

Michael F Cohen and John R Wallace. 1993. *Radiosity and realistic image synthesis*. Elsevier.

Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. 2000. Acquiring the reflectance field of a human face. In *SIGGRAPH*.

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 (2018).

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

Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. 2017. Learning to predict indoor illumination from a single image. *ACM Trans. Graphics* 9, 4 (2017).

Stamatios Georgoulis, Konstantinos Rematas, Tobias Ritschel, Mario Fritz, Tinne Tuytelaars, and Luc Van Gool. 2017. What is around the camera?. In *ICCV*.

Clement Godard, Peter Hedman, Wenbin Li, and Gabriel J Brostow. 2015. Multi-view reconstruction of highly specular surfaces in uncontrolled environments. In *3DV*.

Dan B Goldman, Brian Curless, Aaron Hertzmann, and Steven M Seitz. 2010. Shape and spatially-varying brdfs from photometric stereo. *PAMI* 32, 6 (2010).

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*.

Yannick Hold-Geoffroy, Kalyan Sunkavalli, Sunil Hadap, Emiliano Gambaretto, and Jean-François Lalonde. 2017. Deep Outdoor Illumination Estimation. In *CVPR*.

Z. Hui and A. C. Sankaranarayanan. 2017. Shape and Spatially-Varying Reflectance Estimation from Virtual Exemplars. *PAMI* 39, 10 (2017).

Zhuo Hui, Kalyan Sunkavalli, Joon-Young Lee, Sunil Hadap, Jian Wang, and Aswin C. Sankaranarayanan. 2017. Reflectance capture using univariate sampling of BRDFs. In *ICCV*.

Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. 2017. FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks. In *CVPR*.

Carlo Innamorati, Tobias Ritschel, Tim Weyrich, and Niloy J Mitra. 2017. Decomposing single images for layered photo retouching. 36, 4 (2017).

M. K. Johnson and E. H. Adelson. 2011. Shape estimation in natural illumination. In *CVPR*.

Brian Karis and Epic Games. 2013. Real shading in Unreal Engine 4. *SIGGRAPH 2013 Courses: Physically Based Shading Theory Practice* (2013).

Martin Knecht, Georg Tanzmeister, Christoph Traxler, and Michael Wimmer. 2012. Interactive BRDF Estimation for Mixed-Reality Applications. *WSCG* 20, 1 (2012).

Xiao Li, Yue Dong, Pieter Peers, and Xin Tong. 2017a. Modeling surface appearance from a single photograph using self-augmented convolutional neural networks. *ACM Trans. Graphics* 36, 4 (2017).

Zhengqin Li, Kalyan Sunkavalli, and Manmohan Chandraker. 2018. Materials for Masses: SVBRDF Acquisition with a Single Mobile Phone Image. In *ECCV*.

Z. Li, Z. Xu, R. Ramamoorthi, and M. Chandraker. 2017b. Robust Energy Minimization for BRDF-Invariant Shape from Light Fields. In *CVPR*.

Guilin Liu, Duygu Ceylan, Ersin Yumer, Jimei Yang, and Jyh-Ming Lien. 2017. Material Editing using a Physically Based Rendering Network. *ICCV*.

Julio Marco, Quercus Hernandez, Adolfo Munoz, Yue Dong, Adrian Jarabo, Min H Kim, Xin Tong, and Diego Gutierrez. 2017. DeepToF: off-the-shelf real-time correction of multipath interference in time-of-flight imaging. *ACM Trans. Graphics* 36, 6 (2017).

Stephen R Marschner, Stephen H Westin, Eric PF Lafortune, Kenneth E Torrance, and Donald P Greenberg. 1999. Image-based BRDF measurement including human skin. In *Rendering Techniques*.

Wojciech Matusik, Hanspeter Pfister, Matt Brand, and Leonard McMillan. 2003. A Data-Driven Reflectance Model. *ACM Trans. Graphics* 22, 3 (2003).

Abhimitra Meka, Maxim Maximov, Michael Zollhoefer, Avishek Chatterjee, Hans-Peter Seidel, Christian Richardt, and Christian Theobalt. 2018. LIME: Live Intrinsic Material Estimation. In *CVPR*.

Oliver Nalbach, Elena Arabadzhiyska, Dushyant Mehta, H-P Seidel, and Tobias Ritschel. 2017. Deep shading: convolutional neural networks for screen space shading. *Comput. Graph. Forum* 36, 4 (2017).

Shree K. Nayar, Katsushi Ikeuchi, and Takeo Kanade. 1991. Shape from interreflections. *IJCV* 6, 3 (1991).

Shree K. Nayar, Gurunandan Krishnan, Michael D. Grossberg, and Ramesh Raskar. 2006. Fast Separation of Direct and Global Components of a Scene Using High Frequency Illumination. *ACM Trans. Graphics* 25, 3 (2006).

Diego Nehab, Szymon Rusinkiewicz, James Davis, and Ravi Ramamoorthi. 2005. Efficiently combining positions and normals for precise 3D geometry. In *ACM transactions on graphics (TOG)*, Vol. 24. ACM, 536–543.

Alejandro Newell, Kaiyu Yang, and Jia Deng. 2016. Stacked Hourglass Networks for Human Pose Estimation. In *ECCV*.

Matthew O'Toole and Kiriakos N. Kutulakos. 2010. Optical Computing for Fast Light Transport Analysis. *ACM Trans. Graphics* 29, 6, Article 164 (2010).

Geoffrey Oxholm and Ko Nishino. 2016. Shape and reflectance estimation in the wild. *PAMI* 38, 2 (2016), 376–389.

Ravi Ramamoorthi and Pat Hanrahan. 2001. An efficient representation for irradiance environment maps. In *SIGGRAPH*.

Konstantinos Rematas, Tobias Ritschel, Mario Fritz, Efstratios Gavves, and Tinne Tuytelaars. 2016. Deep reflectance maps. In *CVPR*.

Kosta Ristovski, Vladan Radosavljevic, Slobodan Vucetic, and Zoran Obradovic. 2013. Continuous Conditional Random Fields for Efficient Regression in Large Fully Connected Graphs.. In *AAAI*.

J. Riviere, P. Peers, and A. Ghosh. 2016. Mobile Surface Reflectometry. *Comput. Graph. Forum* 35, 1 (2016).

O. Ronneberger, P.Fischer, and T. Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *MICCAI*.

Soumyadip Sengupta, Angjoo Kanazawa, Carlos D. Castillo, and David W. Jacobs. 2018. SfSNet: Learning Shape, Refectance and Illuminance of Faces in the Wild. In *CVPR*.

Jian Shi, Yue Dong, Hao Su, and Stella X Yu. 2017. Learning Non-Lambertian Object Intrinsics Across ShapeNet Categories. In *CVPR*.

Z. Shu, E. Yumer, S. Hadap, K. Sunkavalli, E. Shechtman, and D. Samaras. 2017. Neural Face Editing with Intrinsic Image Disentangling. In *CVPR*.

A. Tewari, M. Zollhofer, H. Kim, P. Garrido, F. Bernard, P. Perez, and C. Theobalt. 2018. MoFA: Model-Based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. In *ICCV*.

A. Toshev and C. Szegedy. 2014. DeepPose: Human Pose Estimation via Deep Neural Networks. In *CVPR*.

Ting-Chun Wang, Manmohan Chandraker, Alexei Efros, and Ravi Ramamoorthi. 2017. SVBRDF-Invariant Shape and Reflectance Estimation from Light-Field Cameras. *PAMI* (2017).

S. E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. 2016. Convolutional Pose Machines. In *CVPR*.

Robert J. Woodham. 1980. Photometric Method For Determining Surface Orientation From Multiple Images. *Optical Engineering* 19 (1980).

Hongzhi Wu and Kun Zhou. 2015. AppFusion: Interactive Appearance Acquisition Using a Kinect Sensor. *Comput. Graph. Forum* 34, 6 (2015).

Zexiang Xu, Kalyan Sunkavalli, Sunil Hadap, and Ravi Ramamoorthi. 2018. Deep image-based relighting from optimal sparse samples. *ACM Trans. Graphics* 37, 4 (2018).

Yizhou Yu, Paul Debevec, Jitendra Malik, and Tim Hawkins. 1999. Inverse Global Illumination: Recovering Reflectance Models of Real Scenes from Photographs. In *SIGGRAPH*.