# Simultaneous Localization and Appearance Estimation with a Consumer RGB-D Camera

Hongzhi Wu, Zhaotian Wang, and Kun Zhou, *Fellow, IEEE*

**Abstract**—Acquiring general material appearance with hand-held consumer RGB-D cameras is difficult for casual users, due to the inaccuracy in reconstructed camera poses and geometry, as well as the unknown lighting that is coupled with materials in measured color images. To tackle these challenges, we present a novel technique, called Simultaneous Localization and Appearance Estimation (SLAE), for estimating the spatially varying isotropic surface reflectance, solely from color and depth images captured with an RGB-D camera under unknown environment illumination. The core of our approach is a joint optimization, which alternates among solving for plausible camera poses, materials, the environment lighting and normals. To refine camera poses, we exploit the rich spatial and view-dependent variations of materials, treating the object as a localization-self-calibrating model. To recover the unknown lighting, measured color images along with the current estimate of materials are used in a global optimization, efficiently solved by exploiting the sparsity in the wavelet domain. We demonstrate the substantially improved quality of estimated appearance on a variety of daily objects.

**Index Terms**—RGB-D camera, spatially varying BRDF, joint optimization.

## 1 Introduction

With the wide availability of consumer RGB-D cameras, casual users can easily acquire the geometry of various objects at home nowadays. For example, by moving and pointing a Kinect sensor towards an object from different perspectives, a continuous stream of depth maps are obtained; each map is processed, to solve for the corresponding camera pose and refine the current geometry estimate at the same time [1], [2].

The geometry alone is often not sufficient to faithfully convey the realism of a captured object. The key missing piece here is a realistic *material appearance*. Existing systems such as [2] blend multi-view color samples, resulting in ghosting or blurring artifacts. Recently, Zhou and Koltun [3] present an efficient algorithm that jointly estimates RGB-D camera poses and color textures of objects. But it is still difficult for a non-professional user to acquire complex, general material appearance, which varies in space as well as with lighting and view conditions, using an RGB-D camera.

Previous work on material acquisition does an excellent job in digitizing real-world materials with high fidelity [4]. The majority of related work assumes carefully calibrated cameras and precise geometry, and focuses on reconstructing high-dimensional material appearance from measurements. However, a number of challenges arise, if one directly applies existing methods to estimate general material appearance, using a hand-held RGB-D camera at home/office with uncontrolled illumination. First, the camera poses estimated from the noisy depth camera are too inaccurate to recover appearance. Second, the reconstructed geometry, in particular normals, lacks the precision for appearance computation of acceptable quality. Third, unlike many previous approaches that employ active illumination, the unknown lighting must be handled in order to estimate materials from captured color images. Overall, the problem is substantially more complicated, compared with existing work that models color textures [3].

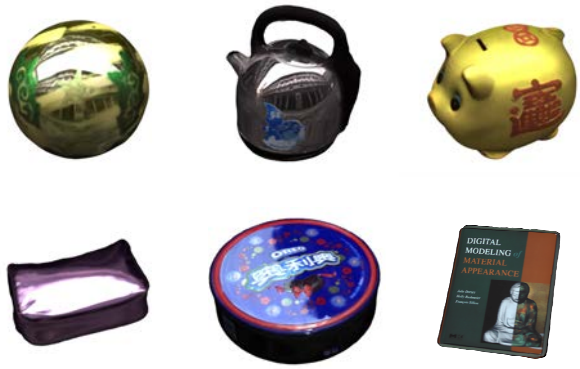


Fig. 1. Rendering results of material appearance estimated using our approach, under the uffizi environment map. Given a geometric model and RGB images acquired by a consumer RGB-D camera, our joint optimization alternates among solving for plausible camera poses, materials, the environment lighting and normals.

In this paper, we present a novel technique for estimating the spatially varying isotropic surface reflectance under unknown environment illumination, solely from color and depth images captured with a hand-held RGB-D camera. To tackle the aforementioned difficulties, we propose a coherent, joint optimization formulation, that alternates among solving for plausible camera poses, materials, the environment lighting and normals. To refine imprecise camera localization, we exploit the rich spatial and view-dependent variations of materials. Essentially, the object is treated as

• *The authors are with State Key Lab of CAD & CG, Zhejiang University, Hangzhou, China, 310058. E-mail: hwu@acm.org, zhaotianzju@gmail.com, kunzhou@acm.org. K. Zhou is the corresponding author.*

*Manuscript received August 19, 2015; revised October 18, 2015.*

a *localization-self-calibrating* model. To recover the unknown lighting, measured color images along with the current estimate of materials are used in a global optimization, which is efficiently solved by exploiting the sparsity in the wavelet domain. We also correct inaccurate normals, mainly based on a photometric consistency constraint.

Our approach considerably simplifies appearance acquisition for casual users at home/office: one only needs to take a video of an object using a hand-held RGB-D camera from different viewpoints, without any markers or explicit calibrations; there is no need to capture the environment illumination using a light probe in a separate process. We test our system on hand-held scans of a variety of daily objects, and demonstrate the substantially improved quality of estimated appearance, ranging from Lambertian to highly specular (Fig. 1). Our system is potentially useful for many applications, such as digital content creation by non-professionals in e-commerce and games.

## 2 Previous Work

To optimize camera poses corresponding to input RGB images is a classic problem in the acquisition of geometric models with color attributes. A number of methods (e.g., [5], [6]) have been proposed to maximize the color consistency among input images. Recently, Zhou and Koltun [3] propose a joint optimization algorithm to compute color textures of Lambertian objects with an RGB-D camera. Previous work typically filters highlights as outliers and produces view-independent texture maps, which are problematic for relighting, as materials and lighting are baked together. In comparison, our method handles more general materials that can be described as 6D SVBRDFs (spatially varying BRDFs), and exploits lighting/view-dependent cues to localize the camera. The SVBRDF, the environment lighting and the normals are also estimated/decoupled as results of our joint optimization.

SVBRDFs can be measured with high precision, by carefully sampling the 6D domain of lighting and view directions, as well as locations. The majority of previous work in this field relies on active illumination to accurately and robustly reconstruct the surface reflectance (e.g., [7], [8], [9]). These methods usually require specific devices together with careful calibrations to obtain full control over the incident lighting, which is not easily accessible to non-professional users. On the other hand, passive appearance acquisition techniques estimate the reflectance with unknown lighting. We review two main classes of passive methods and previous work based on similar hardware, as they are closely related to this paper. Readers are referred to [4] for an excellent survey on recent acquisition techniques.

**Example-based Acquisition.** Hertzmann and Seitz [10] reconstruct normals and reflectance, by photographing an object along with a reference object of known geometry and similar materials. The idea is extended to the multi-view case in [11]. Dong et al. [12] employ a custom-built device for quickly capturing representative BRDFs, and propose a two-pass algorithm that acquires an SVBRDF as linear combinations of the representatives. Ren et al. [13] use a linear light source and a BRDF chart which contains tiles of a variety of known BRDFs. They photograph a planar sample with the chart. The SVBRDF is then reconstructed by aligning the reflectance sequences of the sample and the chart, via dynamic time warping. In comparison, our method does not require the presence of known, reference materials that are similar to the appearance of the object during acquisition. Moreover, our reconstructed SVBRDFs are not limited in the linear subspace spanned by a few example materials.

**Joint Estimation of Reflectance and Lighting.** From a single image, Romeiro and Zickler [14] estimate the homogeneous reflectance on a sphere with unknown lighting, constrained by the statistics of real-world illuminations. For an object of known shape, both the reflectance and lighting can be estimated with constraints on the entropy as well as the bound and variability of real-world materials [15]. The same research group [16] also proposes a method to jointly estimate unknown reflectance and shape, by exploiting the orientation clues in a *known* lighting environment. Haber et al. [17] optimize the lighting and SVBRDFs in an all-frequency wavelet domain based on inverse rendering, while requiring manual registrations of input images. Li et al. [18] reconstruct a geometry using multi-view stereo for human performance, and optimize both the lighting and SVBRDFs, assuming that the specular BRDFs can be divided into spatial clusters of same materials. Palma et al. [19] use video frames and a known geometry as input. They estimate the environment lighting via points with specular reflections, and then optimize a parametric SVBRDF. Recently, Dong et al. [20] reconstruct the lighting and the SVBRDF expressed by a data-driven microfacet model, from a video of a rotating object with known geometry. The sparsity of natural illumination in the gradient domain is exploited to constrain the optimization.

**Similar Hardware Setups.** Recently, researchers start to investigate appearance acquisition using consumer RGB-D cameras. Knecht et al. [21] interactively estimate BRDFs from a fixed-view depth map captured by a Kinect sensor. The surrounding lighting is acquired using a DSLR with a fish-eye lens. In our earlier work [22], we propose a hybrid system that employs the Kinect infra-red emitter/receiver to estimate spatially varying material roughness, and uses the Kinect RGB camera to compute the diffuse and specular albedos. As the focus of the work is to provide quick visual feedback, the camera poses from KinectFusion are used directly. In addition, a separate scanning pass is required to capture the environment illumination from a mirror ball. Zollhöfer et al. [23] refine the geometry captured with an RGB-D camera with shading cues, assuming that surfaces are predominantly Lambertian.

Most existing appearance acquisition methods assume that the camera poses and the object geometry are known and sufficiently accurate (after calibrations), and focus on estimating the materials. However, this is not the case with consumer RGB-D cameras, which motivates our joint optimization framework. We directly take unregistered RGB images from the Kinect sensor plus the inaccurate geometry and camera poses from KinectFusion as input, and explicitly solve for plausible camera poses, materials, the lighting and normals, based on a unified optimization objective.

## 3 PRELIMINARIES

In this section, we derive equations for efficiently computing our joint optimization objective to be defined in Sec. 4. We do not handle visibility information or interreflections in our pipeline, as is common in existing work on appearance acquisition. Although the following derivations are based on grayscale values, the extension to RGB channels is straightforward (Sec. 5). First of all, the outgoing radiance $L$ at a surface point $x$ along a direction $\omega_o$ can be calculated as:

$$L(\omega_o; x) = \int_\Omega E(\omega_i) f_r(\omega_i', \omega_o'; x)(n(x) \cdot \omega_i)^+ d\omega_i. \quad (1)$$

Here $\Omega$ is the upper hemisphere, $E$ is a distant environment lighting, $\omega_i$ is a lighting direction, $f_r$ is a spatially varying isotropic BRDF and $n$ is a normal. We parameterize $E$ over two squares with a total resolution of 256×512, according to [24]: each hemisphere of directions are mapped onto one square (Fig. 2). Note that $\omega'$ denotes a direction in the local frame of $x$, while $\omega$ is expressed in the global coordinate system. In addition, $(\cdot)^+$ is the cosine of the angle between two vectors, which is clamped to zero if negative. We drop $x$ in subsequent derivations for brevity.

Similar to previous work in real-time rendering [25], we represent the 2D cosine-weighted BRDF slice at an outgoing direction $\omega_o'$, as a diffuse term plus a specular term:

$$f_r(\omega_i'; \omega_o') \cdot (n \cdot \omega_i)^+ \approx \frac{\rho_d}{\pi}(n \cdot \omega_i)^+ + \rho_s \alpha \mathcal{G}(\omega_i; \kappa, \mu), \quad (2)$$

where $\rho_d$ is the diffuse albedo, $\rho_s$ is the specular albedo and $\alpha$ is a non-negative scalar. $\mathcal{G}$ is a Gaussian-like von Mises-Fisher (vMF) probability distribution function over the unit sphere [26], defined as:

$$\mathcal{G}(\omega; \kappa, \mu) = \frac{\kappa}{4\pi \sinh(\kappa)} e^{\kappa(\omega \cdot \mu)}, \quad (3)$$

with $\kappa$ as the inverse width and $\mu$ the central direction. Using vMFs allows us to compactly represent a wide range of BRDFs, as only $\{\alpha(\omega_o'), \kappa(\omega_o'), \mu(\omega_o')\}$ for discretized $\omega_o'$ are stored. Moreover, rendering vMF-based BRDFs under environment lighting is highly efficient (Eq. 6). vMF-based BRDFs also lead to a straightforward equation for estimating the environment lighting (Sec. 4.3). On the other hand, the approximation of Eq. 2 has a limited effect on accuracy in our case, which will be detailed in Sec. 6.

With the vMF-based BRDF representation, (inverse) rendering under environment lighting can be performed rapidly. We first precompute the convolution of the lighting $E$ and cosine / vMF lobes as:

$$E_d(n) = \frac{1}{\pi} \int_{S^2} E(\omega_i)(n \cdot \omega_i)^+ d\omega_i, \quad (4)$$

$$E_s(\kappa, \mu) = \int_{S^2} E(\omega_i) \mathcal{G}(\omega_i; \kappa, \mu) d\omega_i, \quad (5)$$

which yields the diffuse response function $E_d(n)$ and the specular response function $E_s(\kappa, \mu)$. An example is shown in Fig. 2. $E_d$ and $E_s(\kappa; \cdot)$ share the same double-squares parameterization as $E$. For $\kappa$, we discretize it as $\kappa = 2, 2^2, ..., 2^{14}$, whose range and sampling rate are sufficient for approximating BRDFs of our interest. Next, by substituting Eq. 4 & 5, computing Eq. 1 is simply a mixture of lookups and basic arithmetic operations:

$$L(\omega_o) \approx \rho_d E_d(n) + \rho_s \alpha(\omega_o') E_s(\kappa(\omega_o'), \mu(\omega_o')). \quad (6)$$

Note that $E_d$, $E_s$ are decoupled from the BRDF $f_r$, and only need to be precomputed once for a given environment lighting $E$. For any $f_r$ expressed in the form of Eq. 2, we can rapidly evaluate the rendering equation (Eq. 6) with precomputed $E_d$ and $E_s$.

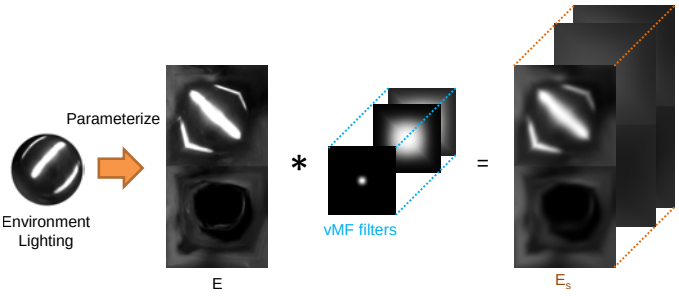



Fig. 2. The environment lighting $E$ is parameterized using an octahedron mapping [24]. The specular response function $E_s$ is computed by convolving $E$ with vMFs of different sizes (i.e., $\kappa$).

## 4 THE JOINT OPTIMIZATION

Our input is a set of images $\{I_j\}$ captured from a Kinect sensor, and a set of vertices $\{x\}$ representing the geometry of an object of interest, obtained from KinectFusion. Our goal is to estimate the appearance of an object that best matches its measurements in $\{I_j\}$. Since the appearance is related to the unknown spatially varying BRDF $f_r$, the environment lighting $E$, the inaccurate camera poses $\{T_j\}$ and normals $\{n\}$ (Eq. 1), we perform a joint optimization with respect to all four factors as follows:

$$\underset{\{T_j\}, \{f_r, n\}_x, E}{\arg\min} \quad \lambda_I P_I + \lambda_S P_S + \lambda_C P_C, \quad (7)$$

where $P_I$ is the photometric consistency term defined as:

$$P_I = \sum_j \sum_{x \in X_j} ||I_j(x, T_j) - L(\omega_o; x, E, T_j)||^2, \quad (8)$$

and $P_S / P_C$ are geometric terms related to the normal optimization only, which will be detailed in Sec. 4.4. In Eq. 8, $X_j$ is the subset of vertices that are visible in image $I_j$. $T_j$ is an extrinsic $4 \times 4$ matrix, that transforms $x$ from the global coordinate system to a local system corresponding to image $I_j$.

Our objective in Eq. 7 is non-convex, and involves many variables. To minimize it in practice, we alternate among solving for camera poses $\{T_j\}$ (Sec. 4.1), materials $f_r$ (Sec. 4.2), the environment lighting $E$ (Sec. 4.3) and normals $n$ (Sec. 4.4). In each stage, we estimate some variables and keep others fixed, while minimizing the same global objective. Pseudo-code of our joint optimization can be found in Tab. 1.

Observe that Eq. 7 is a non-linear least-squares problem, in the form of $\sum_i {r_i}^2$. We can minimize it via the Gauss-Newton method, similar to [3]. Specifically, suppose $\theta$ are

TABLE 1
Pseudo-code of our joint optimization.

| |
|---|
| 1. Initialization (Sec. 5). |
| 2. Solve for camera poses $\{T\}$ (Sec. 4.1). |
| 3. Solve for cluster materials $\{\rho_d, \rho_s, m\}$ (Sec. 4.2). |
| 4. Reassign each point to the closest cluster material (Sec. 4.2). |
| 5. Solve for the lighting $E$ (Sec. 4.3). |
| 6. Solve for normals $\{n\}$ (Sec. 4.4). |
| 7. Go to step 2 if convergence is not reached. |
| 8. Post-processing (Sec. 5). |

the variables of interest at the current stage. In each iteration, $\theta$ is updated as $\theta^{k+1} = \theta^k + \Delta\theta$, where $\Delta\theta$ is the solution to the linear system $J_r^\top J_r \Delta\theta = -J_r^\top r$. Here $r = (r_0, r_1, ...)^\top$ is the residual vector, and $J_r$ is the Jacobian, computed as the finite difference.

### 4.1 Camera Pose Optimization

The camera poses $\{T_j\}$ determine both the measured image pixel $I_j(x, T_j)$ that a point $x$ corresponds to, and the inverse rendering of estimated appearance, $L(\omega_o; x, T_j)$. As $\{T_j\}$ are independent from each other, we optimize them one at a time. Specifically, for an image $I_j$, minimizing Eq. 7 with respect to the transform $T_j$ is equivalent to minimizing the following equation:

$$\arg\min_{T_j} \sum_{x \in X_j} ||I_j(x, T_j) - L(\omega_o; x, T_j)||^2. \tag{9}$$

To obtain $I_j(x, T_j)$, we first compute the transformed point $T_j x$, which is then projected onto the image plane of $I_j$, based on the camera's intrinsic parameters; we then bilinearly interpolate image pixels to get the result. $L$ can be quickly computed with Eq. 6, based on the current estimate of the BRDF, lighting and normal.

Following [3], $\Delta T_j$ is parameterized by a 6D vector:

$$\Delta T_j \approx \begin{pmatrix} 1 & -\gamma_j & \beta_j & a_j \\ \gamma_j & 1 & -\alpha_j & b_j \\ -\beta_j & \alpha_j & 1 & c_j \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{10}$$

where $(a_j, b_j, c_j)^\top$ is a translation, and $(\alpha_j, \beta_j, \gamma_j)^\top$ can be viewed as the angular velocity. In each iteration, we compute $\{\Delta T_j\}$ and update $\{T_j\}$, using the Gauss-Newton method.

### 4.2 Material Optimization

Materials are related to the estimated appearance $L(\omega_o; x)$ via Eq. 1. We describe materials as a 6D SVBRDF, which has a large number of degrees of freedom. To make our material optimization feasible, we assume that there are a finite number of possible specular BRDFs, whose vMF representations are precomputed (Eq. 2). So optimizing $f_r$ is equivalent to optimizing $\rho_d$ and $\rho_s$, and picking the best possible $\{\alpha(\omega_o'), \kappa(\omega_o'), \mu(\omega_o')\}$ from all precomputed specular BRDFs. In this paper, we use 256 isotropic Ward models [27] as specular BRDFs, whose roughness parameter ranges from $0.007$ to $0.4$. Other analytic or measured BRDFs can also be included, provided that they can be well approximated by Eq. 2.

To further constrain the unknown materials, we assume in this stage that each $f_r(x)$ belongs to only one of $k$ distinct BRDFs, where $k$ is a user-specified number. Thus, $X$ is partitioned into $k$ clusters as $X = \cup_l M_l$, based on the BRDFs. This idea is similar to [28].

Now material optimization is performed as two steps in each iteration: computing the optimal cluster BRDFs from measurements of cluster members $M_l$, and reassigning all points to their closest cluster BRDFs. In the first step, we minimize the following goal derived from Eq. 7, for each material cluster $M_l$:

$$\arg\min_{\rho_{d,l}, \rho_{s,l}, m_l} \sum_j \sum_{x \in X_j \cap M_l} ||I_j(x) - L(\omega_o; x)||^2. \tag{11}$$

Here $m_l$ is the index of precomputed specular BRDFs. To find the optimal $m_l$, we first enumerate all precomputed specular BRDFs. For a given $m_l$, $\rho_{d,l}$ and $\rho_{s,l}$ can be calculated as the solution to a non-negative linear least-squares problem, derived by substituting Eq. 6 into Eq. 11:

$$\begin{aligned} &\sum_{j, x \in X_j \cap M_l} ||I_j(x) - L(\omega_o; x)||^2 \\ \approx &\sum_{j, x \in X_j \cap M_l} ||I_j(x) - \rho_{d,l} E_d(n) - \rho_{s,l} \alpha_{m_l} E_s(\kappa_{m_l}, \mu_{m_l})||^2. \end{aligned} \tag{12}$$

The $\{\rho_{d,l}, \rho_{s,l}, m_l\}$ that minimizes Eq. 11 is finally chosen as the optimal cluster material.

The second step in material optimization is to reassign each point $x$ to the closest cluster BRDF, based on its image measurements. For a given $x$, we optimize the following objective:

$$\arg\min_l \sum_j ||I_j(x) - L(\omega_o; x)||^2, \tag{13}$$

by looping over all cluster BRDFs and selecting the one that minimizes the above equation.

### 4.3 Lighting Optimization

Similar to materials, the lighting $E$ is related to the estimated appearance via Eq. 1 as well. We derive the lighting optimization objective from Eq. 7 as:

$$\arg\min_E \sum_j \sum_{x \in X_j} ||I_j(x, T_j) - L(\omega_o; x, E)||^2, \tag{14}$$

$$\approx \sum_j \sum_{x \in X_j} ||I_j(x, T_j) - \int_{S^2} [\frac{\rho_d}{\pi}(n, \omega_i)^+ + \rho_s \alpha \mathcal{G}(\omega_i; \kappa, \mu)] E(\omega_i) d\omega_i||. \tag{15}$$

Note that Eq. 15 is obtained by substituting Eq. 2. As the integral operator is linear, Eq. 15 is essentially a linear least-squares problem. However, solving the above problem in a brute-force way is prohibitively expensive. Consider formulating the equation as $\min ||Fe - p||^2$, where $e = (E(\omega_{i,0}), E(\omega_{i,1}), ...)^\top$, $p = (I_{j_0}(x_0), I_{j_0}(x_1), ..., I_{j_1}(x_0), ...)^\top$, and $F$ is a matrix defined as $F_{kl} = [\frac{\rho_d}{\pi}(n, \omega_{i,l})^+ + \rho_s \alpha \mathcal{G}(\omega_{i,l}; \kappa_k, \mu_k)] \Delta\omega_{i,l}$. Here $\omega_{i,l}$ is a discretization of the lighting direction $\omega_i$. In our experiments, matrix $F$ is large with ~$10^8$ rows and ~$10^5$ columns.

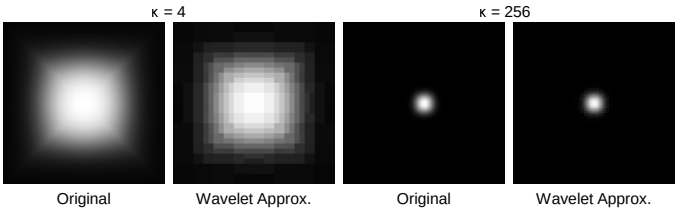



Fig. 3. Two vMFs ($\mathcal{G}$) approximated with Haar wavelets. The number of non-zero coefficients is reduced from 131,070 to 206 ($\kappa$ = 4), and 23,332 to 224 ($\kappa$ = 256), by transforming from the spatial domain to the wavelet domain, while retaining 99.5% of the original energy.

To make the computation of $E$ involving $F$ feasible, we reduce the footprint of $F$ with the following two techniques.

First, we decrease the footprint of each row by exploiting sparsity. We observe that cosine lobes and $\mathcal{G}$ are sparse in the wavelet domain (Fig. 3), owing to the scale-varying basis functions. Note that the same property does not exist in the spatial domain, except for $\mathcal{G}$ with very large $\kappa$. As $E$, $(n,\omega_i)^+$ and $\mathcal{G}$ are parameterized over the double-squares (Fig. 2), we can transform all of them into the Haar wavelet domain. Then Eq. 15 becomes:

$$\arg\min_{E} \sum_{j,x\in X_j} ||I_j(x) - \sum_t E_t[\frac{\rho_d}{\pi}\mathcal{D}_t(n) + \rho_s\alpha\mathcal{G}_t(\kappa,\mu)]||^2,$$

where

$$\begin{aligned} E(\omega_i) &= \sum_t E_t\Psi_t(\omega_i), \\ (n,\omega_i)^+\Delta\omega_i &= \sum_t \mathcal{D}_t(n)\Psi_t(\omega_i), \\ \mathcal{G}(\omega_i;\kappa,\mu)\Delta\omega_i &= \sum_t \mathcal{G}_t(\kappa,\mu)\Psi_t(\omega_i). \end{aligned} \tag{16}$$

Here $E_t$, $\mathcal{D}_t$ and $\mathcal{G}_t$ are wavelet coefficients, and $\Psi_t$ is a basis function. We precompute $\{\mathcal{D}_t\}$ and $\{\mathcal{G}_t\}$ by retaining $99.5\%$ of the total energy.

Second, we approximate $F$ by sampling its rows via [29], to reduce the number of rows. Each original row is sampled with a probability proportional to its $L^2$-norm, efficiently approximated as $(n,\omega_o)^+\sqrt{(\rho_d N_d)^2 + (\rho_s\alpha N_s)^2}$. Here $N_d/N_s$ is the precomputed $L^2$-norm of a Lambertian/vMF lobe. Essentially, each row is weighted by $(n,\omega_o)^+$, to penalize unreliable grazing angle views. We choose the row sampling technique, because it is highly efficient on our large, sparse $F$ in the wavelet domain, and its accuracy is comparable with more computationally involved techniques like random projections, as reported in [30]. In our experiments, row sampling is performed $2\times10^6$ times.

Now we obtain a row-reduced, sparse matrix to approximate $F$ in the wavelet domain. The maximum size of the approximation is only a few GB in our experiments, which not only is manageable in storage, but also allows efficient computation. In comparison, the size of the original dense $F$ is on the order of 100TB. Next, we select corresponding elements of $p$, based on sampled rows of $F$. Finally, we apply a standard iterative solver for sparse least squares [31] to efficiently solve Eq. 16. The results $\{E_t\}$ are then transformed back to the spatial domain to obtain $E$.

**Note.** Due to the band-limiting nature of BRDFs over the environment lighting [32], we are only able to recover $E$ up to the frequency limit represented by the most specular material in the object. Nevertheless, higher frequency content of $E$ is not needed in estimating materials (see the Piggy-Bank example in Fig 14). We describe how to resolve the lighting-material ambiguities in Sec. 5.

### 4.4 Normal Optimization

The normals $\{n(x)\}$ are also related to the estimated appearance via Eq. 1. Similar to existing work such as [33], we constrain the high degrees of freedom in the spatially varying $n(x)$, using a normal smoothness term $P_S$ and a normal integrability term $P_C$, in addition to the photometric consistency term $P_I$ in Eq. 7. We define $P_S$ and $P_C$ as follows:

$$P_S = \sum_x ||n(x) - \frac{\sum_{y\in R_x} n(y)}{||\sum_{y\in R_x} n(y)||}||^2, \tag{17}$$

$$P_C = \sum_x [\frac{1}{\sqrt{A(R_x)}} \sum_{y_0,y_1\in R_x} \int_{y_0}^{y_1} n(y)\cdot\frac{y_1-y_0}{||y_1-y_0||}dy]^2. \tag{18}$$

Eq. 17 essentially computes the mesh Laplacian, where $R_x$ indicates the one-ring neighborhood of $x$. Eq. 18 computes the curl at $x$, where $y0, y1$ are connected vertices in $R_x$, and $A(R_x)$ is the area circumscribed by $R_x$. Note that we use $\sqrt{A}$ instead of $A$ as in the original definition of curl, to make $P_C$ scale-independent for the multi-scale optimization (Sec. 5).

We optimize the normal $n(x)$ on a per-point basis. For each point $x$, our objective is as follows:

$$\begin{aligned} \arg\min_{n(x)}\ & \lambda_I \sum_j ||I_j(x) - L(\omega_o;x)||^2 + \\ & \lambda_S||n(x) - \frac{\sum n(y)}{||\sum n(y)||}||^2 + \\ & \lambda_C[\frac{1}{\sqrt{A(R_x)}} \sum_{y_0,y_1\in R_x} \int_{y_0}^{y_1} n(y)\cdot\frac{y_1-y_0}{||y_1-y_0||}dy]^2, \end{aligned} \tag{19}$$

which is derived from Eq. 7 by substituting Eq. 17 & 18. Following [33], we parameterize a normal $n$ over a 2D vector $(u,v)^\top$ such that $n = (u,v,\sqrt{1-u^2-v^2})^\top$, expressed in a local frame built from a previous estimate of $n$. In each iteration, $\Delta u$ and $\Delta v$ are computed to update the corresponding $n$.

## 5 Implementation Details

**Preprocessing.** We apply [34] to sample points $\{x\}$ over the surfaces of the object, reconstructed with KinectFusion. We also follow [3] to select images that exhibit least blur and are within 10ms to the time stamp of the closest depth map, from which the camera pose is derived. Pixels at grazing angle views ($(n,\omega_o) \leq 0.3$) or depth discontinuities are excluded from processing, due to the unreliability in measurements. Unlike in [3], we do not model lens distortions. Instead, we point the RGB camera so that the object shows up approximately in the center region of captured images during acquisition. We find that the distortions in this region do not cause problems for our algorithm.

**Initialization.** As our joint optimization is solved in an iterative fashion, it is important to set good initial values for the unknowns, for the quality of the final results. Specifically, we initialize camera poses $\{T_j\}$ as the ones obtained from KinectFusion after shifting 2.5cm along the local $x$ axis. The shifting is applied, because we need to estimate the RGB camera poses from the depth camera poses obtained with KinectFusion, where the baseline between two cameras is 2.5cm. Initial material clusters are generated based on the initial diffuse albedo $\tilde{\rho_d}(x)$, computed as: $\tilde{\rho_d}(x) = \min_j I_j(x)$, assuming that $E_d(n) = 1$. Next, we initialize the environment lighting $E$, by assuming that each $f_r$ has a diffuse albedo of $\tilde{\rho_d}(x)$ and a perfect mirror specular BRDF with a unit specular albedo. Then $E$ is computed as $E(2(\omega_o \cdot n)n - \omega_o) = I_j(x) - \tilde{\rho_d}(x)$, where $2(\omega_o \cdot n)n - \omega_o$ is the reflection vector of $\omega_o$ with respect to $n$. With the initial clustering and initial $E$, we calculate the corresponding cluster materials. Moreover, we smooth the normals obtained from KinectFusion as the initial $\{n\}$, using the algorithm described in [35].

**Multi-scale Optimization.** We build a hierarchy of $\{I\}$, $\{E\}$ and $\{x\}$ in a bottom-up fashion. In each level, the resolutions of $\{I\}$ and $\{E\}$ are reduced by a factor of 2, and the number of points $\{x\}$ by a factor of 4. We perform the joint optimization starting from the top level in the hierarchy, and do not move to the next level until convergence. Three levels are used in our experiments.

**Handling Material Boundaries.** Physical material boundaries sometimes result in aliased edges on measured images $\{I\}$, which causes inaccuracy when computing the environment lighting from images in Eq. 15. For these cases, we detect whether a point $x$ is within a certain distance of a cluster boundary, and do not compute the lighting from the image measurements of that point if the former condition holds.

**Post-processing.** After the joint optimization, we compute RGB versions of $\rho_d$ and $\rho_s$ for cluster materials, using the original RGB image $\{I\}$ as input. Next, to refine the reconstructed appearance, we follow a simple method in [28], which projects the image measurements at each point $x$ to the cluster materials, by solving a non-negative least squares problem. The final $f_r(x)$ is represented as a linear combination of cluster materials. We would like to emphasize that our framework is not tied to the projection method. In fact, since we already have plausible estimates of both $\{T_j\}$, $E$ and $\{n\}$, any conventional appearance reconstruction method, such as [36], could be plugged in this stage.

**Lighting-material Ambiguities.** As the lighting and the materials are both unknowns in our optimization, we need to resolve their ambiguities by imposing additional constraints. Since both $E$ and $f_r$ are linear with respect to Eq. 1, exactly the same inverse rendering can be produced, if we scale $E$ by a scalar $a$, and $\rho_d$ and $\rho_s$ by $\frac{1}{a}$. To resolve this scale ambiguity, we normalize $E$ after the lighting optimization, by scaling it such that the average $E$ equals a user-specified $E_{avg}$ ($E_{avg}$ = 1 in our experiments); all $\rho_d$ and $\rho_s$ are adjusted accordingly to retain the same inverse rendering result.

Another fundamental ambiguity is that the angular sharpness of a BRDF can be traded by the blurriness of the lighting [32]. We exploit the statistics of common home/office lighting to determine the absolute BRDF roughness. First, the environment maps in two different lighting conditions are captured from a mirror ball using our pipeline. Next, we compute the specular responses $E_s$ of the average environment map, and analyze their power spectra in the Haar wavelet domain. We find that the normalized histogram of band 4 and 5 is discriminating with respect to $\kappa$. Therefore, we precompute the histograms as the references on $E_s(\kappa;\cdot)$ of various $\kappa$. During runtime, for a particular material, we compute the normalized histogram of band 4 and 5 of $E_s(\hat{\kappa};\cdot)$, where $\hat{\kappa}$ is the precomputed, average $\kappa$ in the vMF representation of the current BRDF. Finally, we find the closest match to the computed histogram in the references, and determine the absolute roughness based on the $\kappa$ corresponding to the match. The above algorithm is executed right before post-processing, and works well in our experiments.

Note that the reference environment maps only need to be acquired once, and are discarded after the analysis. The user is not required to capture any environment map. It will be interesting future work to analyze a larger database of environment maps, similar to [37]. Other methods for determining the absolute roughness can also be employed here.

## 6 Results

All experiments are conducted on a workstation with an Intel i7-4790k CPU and 32GB of memory. The RGB images are captured using a first-generation Kinect sensor at 12fps and a resolution of 1280×960, with fixed exposure and white balancing in a linear space. In precomputing vMF representations of isotropic Ward models, we discretize $\omega_o$ as 180 different elevation angles (i.e., the sampling interval = $0.5°$), which results in only 0.9MB for all 256 BRDFs. The precomputed $\{\mathcal{D}_t, \mathcal{G}_t\}$ (Sec. 4.3) takes up 117MB, 546MB and 3.04GB for three levels in our multi-scale optimization.

To evaluate the approximation accuracy of the vMF-based BRDFs, we compute the relative root-mean-squared error (RMSE) according to [38] for all 256 precomputed BRDFs. We do not consider $\omega_o$ that is below the grazing angle threshold in Sec. 5, as measurements along these directions are excluded from our pipeline. The relative RMSEs for all precomputed BRDFs are in the range of $[0.146, 0.190]$, which are sufficient for our applications.

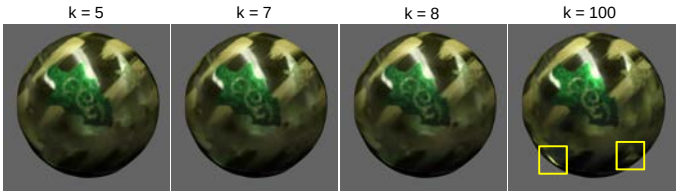



Fig. 5. Impact of using different $k$. Setting $k$ roughly to the number of distinct materials produces plausible results (the left three images), while using a large value (rightmost) makes our optimization under-constrained.

Our objective function in Eq. 7 is complex. There is no theoretical guarantee that our optimization converges to the global optimal solution. Nevertheless, we observe in all

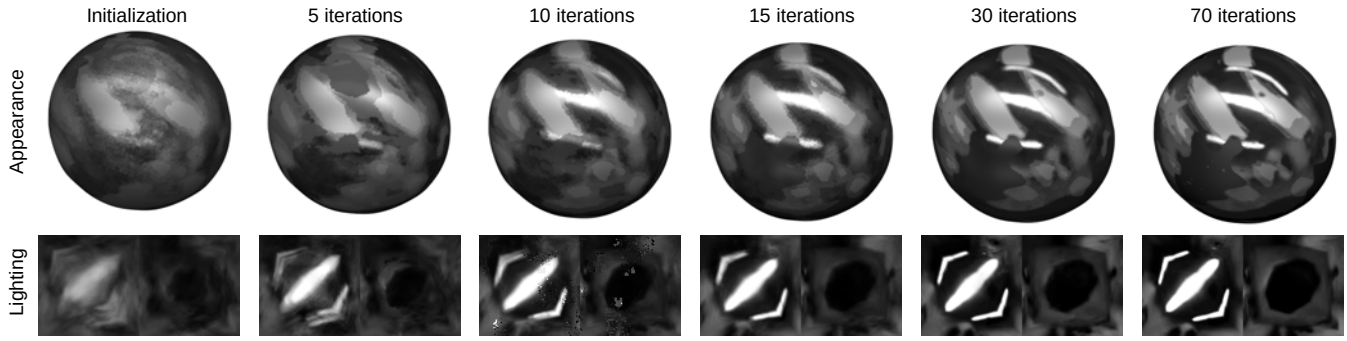



Fig. 4. Progress of our joint optimization on the Xmas-Ball model. The first row shows the estimated appearance rendered with a corresponding estimated environment lighting, which is visualized in the second row. The reference photograph of the model can be found in Fig. 13. No post-processing is performed at this stage.

experiments that the joint optimization converges quickly (typically after 75 iterations), and the objective function decreases over iterations, as illustrated in Fig. 6. While it is possible that the optimization gets trapped in a local minimum, as is common in related work (e.g., [17], [20]), we find in practice that the results of the optimization are sufficiently good as plausible estimates of appearance/normals, which are useful for realistic rendering/editing. In addition, the initial values for camera poses and normals, obtained from KinectFusion, are not drastically different from the ground-truth. Also note that the albedos and the environment lighting are computed using linear least squares, which yields the global minimum.

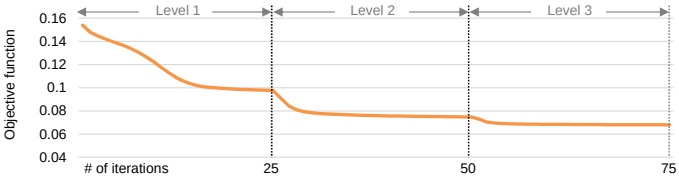



Fig. 6. The optimization objective (Eq. 7) plotted as a function of the number of iterations for the Xmas-Ball case. Three levels are used in the optimization.

The joint optimization on average takes about 100 minutes to compute, with roughly 25%, 20%, 50% and 5% of the time spent on camera pose, material, environment lighting and normal optimizations, respectively. We find that setting $k$ roughly to the number of distinct materials works well in our experiments. As shown in Fig. 5, the final results of our optimization are not sensitive to the choice of $k$. Typical values for $\lambda$ are $\lambda_S = 6$ and $\lambda_I = \lambda_C = 1$. Detailed statistics of our experiments can be found in Tab. 2.

TABLE 2
Statistics of our experiments.

| Model | # of images | # of points | $k$ |
|---|---|---|---|
| Xmas-Ball | 128 | 600K | 6 |
| Kettle | 276 | 1M | 3 |
| Piggy-Bank | 227 | 600K | 4 |
| Cookie-Tin | 194 | 800K | 4 |
| Cosmetic-Bag | 68 | 800K | 1 |
| Book | 32 | 600K | 5 |

The progress of a joint optimization is visualized in Fig. 4. It is interesting to note that, although our objective is to match image measurements, the environment lighting also gets refined indirectly, as a consequence of more accurate estimates of camera poses, normals and materials.

We evaluate the efficacy of our joint optimization with previous work on a variety of objects, whose materials range from diffuse, glossy to highly specular (Fig. 13). Our results are compared with the textured models produced by [3], the results using our optimization with camera poses from KinectFusion, the results using our optimization with camera poses computed by [3] and corresponding photographs. The photographs in Fig. 13 participate in camera pose optimization only, and are excluded from the rest of the joint optimization. In cases where our method is used with camera poses from other approaches, we disable camera pose computation in the joint optimization, and keep the material and environment lighting optimizations. As shown in the figure, the camera poses obtained from KinectFusion are too imprecise for estimating SVBRDFs, particularly specular ones. Note that in the Xmas-Ball case, view-dependent highlights are aligned as view-independent texture maps by [3], which results in large camera pose errors. Moreover, in the highly-textured Cookie-Tin case, accurately-aligned textures can be obtained from [3], despite the white stripes caused by the averaging of highlights. Using our method in conjunction with camera poses from [3] generates a less satisfactory result as well, since such camera poses are optimized without considering the lighting or the SVBRDF. Our method produces a plausible result even for the Piggy-Bank with no specular materials, while other methods fail to align the red coin feature on the back of the pig.

We also compare our method with [3] on a diffuse book in Fig. 7. While the appearance results of [3] are textures only, our method estimates both SVBRDF and lighting separately. Note that the original method in [3] gets trapped in a wrong local minimum. We extend their method with the multi-scale idea in Sec. 5 to obtain a plausible result.

Relighting results of our estimated appearance using the *uffizi* environment map, along with additional details, are shown in Fig. 14. We also demonstrate the environment illumination recovered by our optimization. All objects are captured in an office with light tubes over the ceiling. We do not perform reprojection for specular materials of Kettle, as they are considerably distinct. More results rendered with different conditions can be found in the accompanying video.


Fig. 7. Comparisons between our method and previous work on the diffuse Book model. We estimate both the SVBRDF and lighting, while [3] computes textures only. Note that the top-left photograph is not used in material/lighting/normal optimization.

Repeatability experiments are conducted to test our joint optimization under different illumination and camera trajectories. We capture the same Xmas-Ball model in two separate scans with different lighting conditions, as illustrated in Fig. 8. Relighting results of the estimated appearance are demonstrated as well.

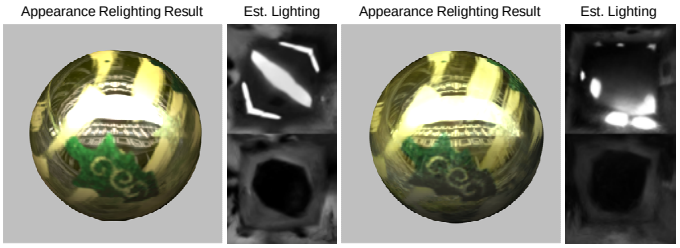



Fig. 8. Results of the same Xmas-Ball model captured under different conditions. For each pair of images, we show the relighting result of the estimated appearance under the uffizi environment map (left), and the illumination recovered by our optimization (right).

## 7 Discussions

We evaluate the impact of each of the four variables in the joint optimization: camera poses, lighting, materials and geometry. Physical and/or synthetic experiments are conducted to study the impact of each variable in isolation.

**Camera Poses.** We evaluate the numerical accuracy of camera positions estimated with various methods. First, we put a checker-board pattern below the Xmas-Ball during acquisition, and reconstruct the reference camera positions using a standard method [39]. Next, we compute the RMSE of camera positions estimated with our approach, KinectFusion and [3], which are 0.071, 0.074 and 0.43, respectively. The numerical error of our method is slightly smaller than that of KinectFusion, but the difference in visual quality of appearance reconstruction is substantial in Fig. 13. The main reason is that specular reflections are highly sensitive to camera poses.

To evaluate the sensitivity of our approach to the error in the camera trajectory, we add unbiased Gaussian noise to the camera poses obtained from KinectFusion, similar to [3], and then perform the joint optimization. Specifically,

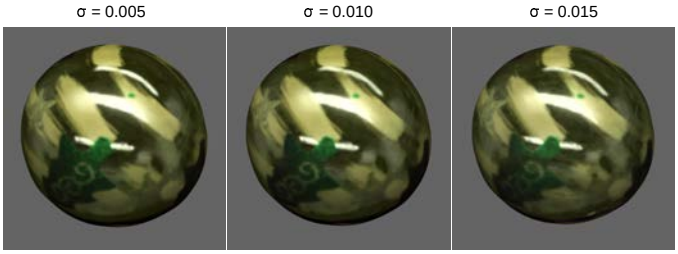



Fig. 9. Impact of camera trajectory perturbations. Gaussian noise with a deviation of $\sigma$ is added to the initial camera poses computed from KinectFusion.

we modify each existing camera pose with an incremental transformation $\Delta T_j(a_j, b_j, c_j, \alpha_j, \beta_j, \gamma_j)$. The translational component $(a_j, b_j, c_j)^\top$ is a direction sampled uniformly at random, then multiplied by a scalar sampled from a Gaussian distribution with a deviation of $\sigma$. The rotational component $(\alpha_j, \beta_j, \gamma_j)^\top$ is computed in the same way. In Fig. 9, we show the robustness of our method when $\sigma = 0.005, 0.010$ and $0.015$ (meter).

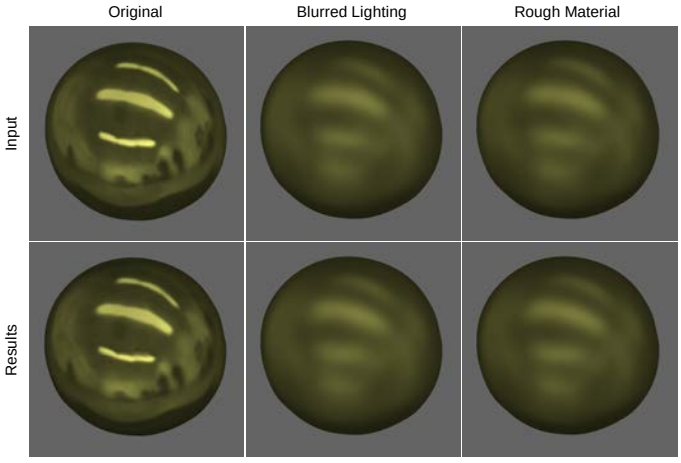



Fig. 10. Impact of a blurred environment lighting / a rough material on optimization results. The blurred environment lighting is generated by convolving with a vMF kernel ($\kappa = 32$). The rough material has a roughness of $0.13$, while the original roughness is $0.007$.

**Lighting.** The presence of a moving person in our acquisition setup makes the environment illumination non-constant. To alleviate this issue, we plan our motion path to avoid blocking of main light sources, which are typically ceiling lights at home/office. In addition, when scanning objects with materials of high specular albedos, we try to stay at a distance to reduce the sizes of the reflections of the person and the Kinect sensor on measured images. In practice, we do not find the issue a problem, as shown in Fig. 13. It is worth noting that the persons captured in the photograph of the Kettle model in Fig. 13 are averaged out and do not show up in our appearance result.

To evaluate the impact of blurred environment lighting, we first conduct a synthetic experiment by rendering the acquired geometry of Xmas-Ball with a highly specular material (roughness = $0.007$) under a captured environment lighting, using camera poses optimized by our method (the left column of Fig 10). The rendered images, along with the camera poses obtained from KinectFusion, are used as input

for our joint optimization. The RMSE for estimated camera positions is $0.7\times10^{-4}$, and the average normal error is $0.43°$. Next, we blur the lighting with a vMF kernel ($\kappa = 32$) and run our optimization again (the center column of Fig 10). The camera position RMSE is $1.6 \times 10^{-4}$, and the normal error is $0.47°$. A visual comparison on estimated appearance can be found in Fig 10.

**Materials.** Similar to the synthetic experiment on blurred lighting, we render the geometry of Xmas-Ball with a rough material (roughness = $0.13$) as input images, and then perform our optimization. Please refer to the right column of Fig. 10 for a visual comparison between the ground-truth and our result. The localization RMSE is $1.9 \times 10^{-4}$, and the normal error is $0.46°$.

**Geometry.** We evaluate the sensitivity of our optimization to the accuracy of the input geometric model on a simplified mesh. As shown in Fig. 11, our approach is robust with respect to geometric model error. In addition, to demonstrate the effect of normal optimization, we show in the same figure a result generated with normals unchanged.


Fig. 11. Impact of geometric model error. We perform our joint optimization on the simplified geometry with (center image) and without the normal optimization (right image).

**LDR Input Images.** The RGB images captured by the Kinect sensor are limited in its dynamic range, compared to previous work that uses DSLRs with bracketing. To evaluate the impact of the low dynamic range, we compare two experiments that use the same set of RGB images as input. The only difference is that one set is of HDR, and the other LDR (Fig. 12). The camera position RMSE and normal error for the HDR case is the same as the original experiment in Fig. 10. For the LDR case, the location RMSE is $0.8 \times 10^{-4}$, and the normal error is $0.42°$, which are similar to the HDR case.

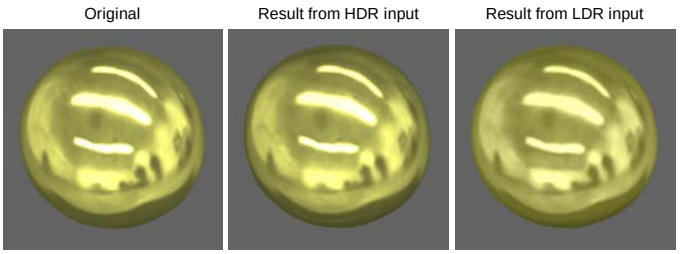



Fig. 12. Impact of LDR / HDR input images on optimization results.

## 8 Limitations and Future Work

Our technique is subject to a number of limitations. First, ignoring occlusions and interreflections will result in less accurate estimate of materials in regions where such effects are strong. Next, the distant lighting assumption restricts the maximum size of object, as the light-object distance in a common home/office is limited. This could be solved by explicitly modeling local lighting effects. Due to the sensor limitation, all RGB input images are captured in low dynamic range and with a low spatial resolution, compared to a DSLR with bracketing. We expect that the quality of our results can get improved, with future hardware updates. In addition, the assumption of a few BRDF clusters is not valid in cases where the object has a large number of distinct BRDFs. Moreover, comparing with latest work on capturing normal maps of predominantly Lambertian objects using consumer RGB-D cameras [23], our resulting normal maps appear over-smoothed and less accurate.

In the future, we would like to extend our framework to handle more general cases, such as objects with anisotropic materials. It will also be interesting to further refine vertex positions of the geometry reconstructed from KinectFusion, using our optimized normals via [40]. As more degrees of freedom are introduced with the changing geometry, proper extra constraints are needed in the joint optimization to avoid getting trapped in undesired local minimums.

## Acknowledgments

The authors would like to thank Lu Zhao and Minmin Lin for proofreading, and Xuchao Gong for help on the video. This work is partially supported by NSF China (No. 61272305 and No. 61303135), and National Program for Special Support of Eminent Professionals of China.

## References

[1] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohli, J. Shotton, S. Hodges, and A. Fitzgibbon, "KinectFusion: Real-time dense surface mapping and tracking," in *Proc. of ISMAR*, 2011, pp. 127–136.

[2] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, and A. Fitzgibbon, "KinectFusion: Real-time 3D reconstruction and interaction using a moving depth camera," in *Proc. of UIST*, 2011, pp. 559–568.

[3] Q.-Y. Zhou and V. Koltun, "Color map optimization for 3D reconstruction with consumer depth cameras," *ACM Trans. Graph.*, vol. 33, no. 4, pp. 1–10, 2014.

[4] T. Weyrich, J. Lawrence, H. P. A. Lensch, S. Rusinkiewicz, and T. Zickler, "Principles of appearance acquisition and representation," *Found. Trends. Comput. Graph. Vis.*, vol. 4, no. 2, pp. 75–191, 2009.

[5] F. Bernardini, I. M. Martin, and H. Rushmeier, "High-quality texture reconstruction from multiple scans," *IEEE Trans. Vis. Comp. Graph.*, vol. 7, no. 4, pp. 318–332, 2001.

[6] M. Corsini, M. Dellepiane, F. Ganovelli, R. Gherardi, A. Fusiello, and R. Scopigno, "Fully automatic registration of image sets on approximate geometry," *IJCV*, vol. 102, no. 1-3, pp. 91–111, March 2013.

[7] J. Lawrence, A. Ben-Artzi, C. DeCoro, W. Matusik, H. Pfister, R. Ramamoorthi, and S. Rusinkiewicz, "Inverse shade trees for non-parametric material representation and editing," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 735–745, 2006.

[8] M. Aittala, T. Weyrich, and J. Lehtinen, "Practical SVBRDF capture in the frequency domain," *ACM Trans. Graph.*, vol. 32, no. 4, pp. 110:1–110:12, Jul. 2013.

[9] B. Tunwattanapong, G. Fyffe, P. Graham, J. Busch, X. Yu, A. Ghosh, and P. Debevec, "Acquiring reflectance and shape from continuous spherical harmonic illumination," *ACM Trans. Graph.*, vol. 32, no. 4, pp. 109:1–109:12, Jul. 2013.


Fig. 13. Comparisons of results using various methods. From the left column to the right: a color-textured model reconstructed by [3], using our method with camera poses obtained from KinectFusion, using our method with camera poses computed by [3], our result and the corresponding photograph (not used in material/lighting/normal optimization).

[10] A. Hertzmann and S. M. Seitz, "Shape and materials by example: A photometric stereo approach," in *Proc. of CVPR*, 2003, pp. 533–540.

[11] A. Treuille, A. Hertzmann, and S. Seitz, "Example-based stereo with general brdfs," in *Proc. of ECCV*, vol. 3022, 2004, pp. 457–469.

[12] Y. Dong, J. Wang, X. Tong, J. Snyder, Y. Lan, M. Ben-Ezra, and B. Guo, "Manifold bootstrapping for SVBRDF capture," *ACM Trans. Graph.*, vol. 29, no. 4, pp. 98:1–98:10, Jul. 2010.

[13] P. Ren, J. Wang, J. Snyder, X. Tong, and B. Guo, "Pocket reflectometry," *ACM Trans. Graph.*, vol. 30, no. 4, pp. 1–10, 2011.

[14] F. Romeiro and T. Zickler, "Blind reflectometry," in *Proc. of ECCV*, vol. 6311, 2010, pp. 45–58.

[15] S. Lombardi and K. Nishino, "Reflectance and natural illumination from a single image," in *Proc. of ECCV*, vol. 7577, 2012, pp. 582–595.

[16] G. Oxholm and K. Nishino, "Shape and reflectance from natural illumination," in *Proc. of ECCV*, vol. 7572, 2012, pp. 528–541.

[17] T. Haber, C. Fuchs, P. Bekaer, H. P. Seidel, M. Goesele, and H. P. A. Lensch, "Relighting objects from image collections," in *Proc. of CVPR*, 2009, pp. 627–634.

[18] G. Li, C. Wu, C. Stoll, Y. Liu, K. Varanasi, Q. Dai, and C. Theobalt, "Capturing relightable human performances under general uncontrolled illumination," *Computer Graphics Forum*, vol. 32, no. 2pt3, pp. 275–284, 2013.

[19] G. Palma, M. Callieri, M. Dellepiane, and R. Scopigno, "A statistical method for SVBRDF approximation from video sequences in general lighting conditions," *Comp. Graph. Forum*, vol. 31, no. 4, pp. 1491–1500, 2012.

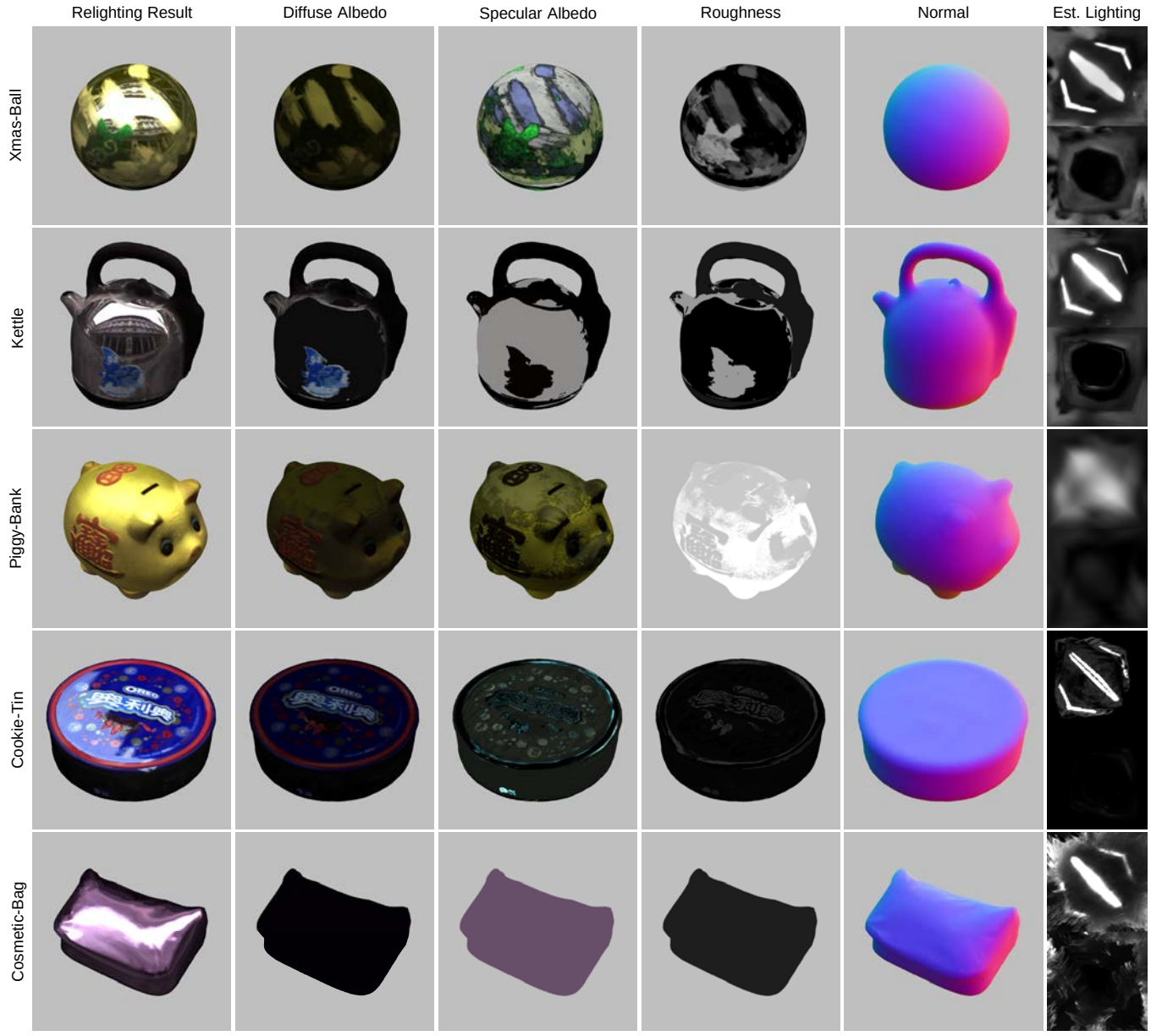



Fig. 14. Appearance and environment lighting results. From the left column to the right: a relighting result of estimated appearance using the uffizi environment map, visualizations of the diffuse albedo, the specular albedo, the roughness and normals, and the illumination estimated by our optimization.

[20] Y. Dong, G. Chen, P. Peers, J. Zhang, and X. Tong, "Appearance-from-motion: Recovering spatially varying surface reflectance under unknown lighting," *ACM Trans. Graph.*, vol. 33, no. 6, pp. 193:1–193:12, Nov. 2014.

[21] M. Knecht, G. Tanzmeister, C. Traxler, and M. Wimmer, "Interactive BRDF estimation for mixed-reality applications," *Journal of WSCG*, vol. 20, no. 1, pp. 47–56, Jun. 2012.

[22] H. Wu and K. Zhou, "AppFusion: Interactive appearance acquisition using a Kinect sensor," *Computer Graphics Forum*, vol. 34, no. 6, pp. 289–298, 2015.

[23] M. Zollhöfer, A. Dai, M. Innmann, C. Wu, M. Stamminger, C. Theobalt, and M. Nießner, "Shading-based refinement on volumetric signed distance functions," *ACM Trans. Graph.*, vol. 34, no. 4, pp. 96:1–96:14, Jul. 2015.

[24] E. Praun and H. Hoppe, "Spherical parametrization and remeshing," *ACM Trans. Graph.*, vol. 22, no. 3, pp. 340–349, Jul. 2003.

[25] P. Green, J. Kautz, W. Matusik, and F. Durand, "View-dependent precomputed light transport using nonlinear Gaussian function approximations," in *Proc. of I3D*, 2006, pp. 7–14.

[26] C. Han, B. Sun, R. Ramamoorthi, and E. Grinspun, "Frequency domain normal map filtering," *ACM Trans. Graph.*, vol. 26, no. 3, pp. 28:1–28:11, July 2007.

[27] J. Dorsey, H. Rushmeier, and F. Sillion, *Digital Modeling of Material Appearance*. Morgan Kaufmann Publishers Inc., 2007.

[28] H. P. A. Lensch, J. Kautz, M. Goesele, W. Heidrich, and H.-P. Seidel, "Image-based reconstruction of spatial appearance and geometric detail," *ACM Trans. Graph.*, vol. 22, no. 2, pp. 234–257, Apr. 2003.

[29] P. Drineas and R. Kannan, "Pass efficient algorithms for approximating large matrices," in *SODA*, vol. 3, 2003, pp. 223–232.

[30] E. Liberty, "Simple and deterministic matrix sketching," in *Proc. of SIGKDD*, 2013, pp. 581–588.

[31] C. C. Paige and M. A. Saunders, "LSQR: An algorithm for sparse linear equations and sparse least squares," *ACM Trans. Math. Softw.*, vol. 8, no. 1, pp. 43–71, Mar. 1982.

[32] R. Ramamoorthi and P. Hanrahan, "A signal-processing framework for inverse rendering," in *Proc. of SIGGRAPH '01*. New York, NY, USA: ACM, 2001, pp. 117–128.

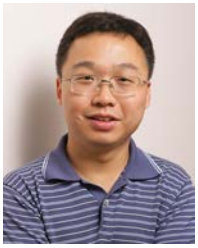
[33] M. K. Johnson and E. H. Adelson, "Shape estimation in natural illumination," in *Proc. of CVPR*, 2011, pp. 2553–2560.

[34] G. Turk, "Re-tiling polygonal surfaces," in *Proc. of SIGGRAPH '92*,

1992, pp. 55–64.
[35] S. Fleishman, I. Drori, and D. Cohen-Or, "Bilateral mesh denoising," *ACM Trans. Graph.*, vol. 22, no. 3, pp. 950–953, Jul. 2003.
[36] R. P. Weistroffer, K. R. Walcott, G. Humphreys, and J. Lawrence, "Efficient basis decomposition for scattered reflectance data," in *Proc. of EGSR*, 2007, pp. 207–218.
[37] R. O. Dror, A. S. Willsky, and E. H. Adelson, "Statistical characterization of real-world illumination," *Journal of Vision*, vol. 4, no. 9, p. 11, 2004.
[38] F. Romeiro, Y. Vasilyev, and T. Zickler, "Passive reflectometry," in *Proc. of ECCV*, vol. 5305, 2008, pp. 859–872.
[39] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. PAMI*, vol. 22, no. 11, pp. 1330–1334, 2000.
[40] D. Nehab, S. Rusinkiewicz, J. Davis, and R. Ramamoorthi, "Efficiently combining positions and normals for precise 3D geometry," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 536–543, Jul. 2005.

**Hongzhi Wu** is currently an assistant professor in State Key Lab of CAD & CG, Zhejiang University. He received B.Sc. in computer science from Fudan University in 2006, and Ph.D. in computer science from Yale University in 2012. His research interests include appearance modeling, design and rendering. He has served on the program committees of PG, EGSR and SCCG.

**Zhaotian Wang** is a master student at State Key Lab of CAD & CG, Zhejiang University. He received his B.Sc. from the same university in 2013. His research interests include appearance acquisition and rendering.



**Kun Zhou** is a Cheung Kong Professor in the Computer Science Department of Zhejiang University, and the Director of the State Key Lab of CAD&CG. Prior to joining Zhejiang University in 2008, Dr. Zhou was a Leader Researcher of the Internet Graphics Group at Microsoft Research Asia. He received his B.S. degree and Ph.D. degree in computer science from Zhejiang University in 1997 and 2002, respectively. His research interests are in visual computing, parallel computing, human computer interaction, and virtual reality. He currently serves on the editorial/advisory boards of ACM Transactions on Graphics and IEEE Spectrum. He is a Fellow of IEEE.