Eurographics Symposium on Rendering - Experimental Ideas & Implementations (2018)
T. Hachisuka and W. Jakob (Editors)

Approximate svBRDF Estimation From Mobile Phone Video

Rachel A. Albert[1,2], Dorian Yao Chan[2], Dan B. Goldman[3], and James F. O'Brien[2]

[1]NVIDIA Research, Santa Clara, CA, USA
[2]The University of California Berkeley, Berkeley, CA, USA
[3]Google, Inc., Seattle, WA, USA

Abstract

We describe a new technique for obtaining a spatially varying BRDF (svBRDF) of a flat object using printed fiducial markers and a cell phone capable of continuous flash video. Our homography-based video frame alignment method does not require the fiducial markers to be visible in every frame, thereby enabling us to capture larger areas at a closer distance and higher resolution than in previous work. Pixels in the resulting panorama are fit with a BRDF based on a recursive subdivision algorithm, utilizing all the light and view positions obtained from the video. We show the versatility of our method by capturing a variety of materials with both one and two camera input streams and rendering our results on 3D objects under complex illumination.

CCS Concepts
•***Computing methodologies*** → ***Reflectance modeling;*** *Computational photography; Texturing;*

1. Introduction

Artistic expression in the creation of virtual objects has grown tremendously in recent years thanks to research in computer generated geometry, lighting, and materials. However, many real-world surfaces exhibit irregular variations in texture and reflectance that are unique and difficult to reproduce algorithmically. Examples include organic materials such as specific pieces of wood or granite, hand-made surfaces such as paintings, and well-worn objects with particular patterns of dirt, scratches, and aging effects. High-quality results can be achieved when these missing details are filled in manually by artists, but doing so requires significant expertise, well-sourced input images, and hours of manual adjustment. Alternatively, it is also possible to obtain high-quality materials through direct capture, but the capture process is also cumbersome due to the specialized equipment typically required.

There are several ways to represent opaque surface reflectances using data derived from the real world. The most common examples are artist-designed materials, direct measurements of real objects, and parametric reflectance models. Artist-designed materials are represented by a set of layers derived from images, wherein each layer describes a component of the reflectance such as the diffuse color, specular behavior, or normal displacement. The process for creating these materials typically involves sourcing a high-quality photograph of a nearly-flat object, and then recombining filtered versions of the photo with procedurally-generated noise layers [Ble13, Ble05]. To obtain a realistic result, artists must expend significant time tweaking parameters via trial-and-error. Libraries of materials, called "material packs" are also widely available for purchase [Pri16], demonstrating both the value of using realistic material models and the effort required to create them. Hand-designed material models generally do not accurately capture the actual reflectance behavior of the real-world material that they are based on. Rather they mimic the real material's appearance, which is sufficient for many rendering applications.

The most complex and complete representations of real materials come from direct measurement. The surface appearance of the object is measured over a densely sampled hemisphere of light and view angles, using a device such as a gonioreflectometer, and these data are interpolated at render time from a lookup table [Mat03]. The measurements span a six dimensional space — azimuth and elevation angles for both the camera and light source and 2D coordinates on the surface — called a spatially varying bi-directional reflectance distribution function (svBRDF) [NRH*77]. Obtaining a measured svBRDF is a time-consuming and memory-intensive process that requires a sample to be brought into a lab with controlled lighting and specialized equipment. Not only are there very few measured svBRDFs available, but this high level of physical accuracy is also generally excessive when only visually plausible images are required.

In many cases the physical plausibility of the material is important, but the reflectance behavior is simple enough it can be accurately represented by a parametric model with only a few parameters. In this case a parametric BRDF model can be created either by choosing arbitrary parameter values, navigating the space of BRDFs with a user interface, or fitting a model to observations of a real object. Well-designed BRDF models adhere to physical

DOI: 10.2312/sre.20181168

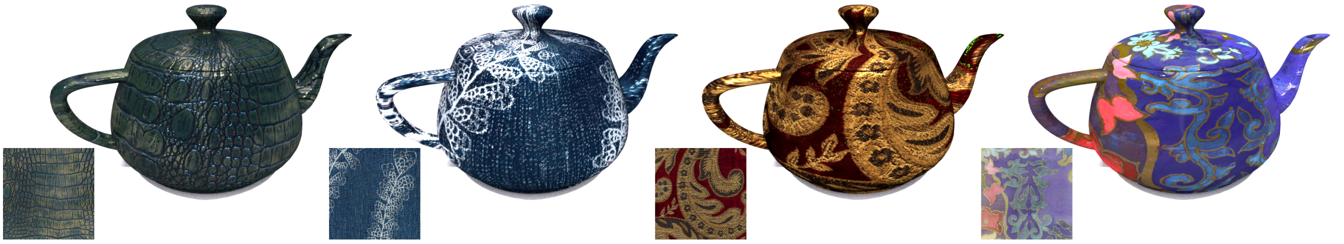


Figure 1: Four materials captured using our method and rendered with Mitsuba. Objects are illuminated by the Uffizi environment map (courtesy of the USC Vision & Graphics Lab) along with two point light sources. The left two materials were captured with one camera, and the two on the right were captured with two cameras. From left to right: green faux leather, blue damask fabric, red velvet lamé, and wrapping paper. Inset image shows the average color per pixel as described in section 3.2.1.

limitations such as conservation of energy and reciprocity, without requiring the significant memory overhead of a measured BRDF.

We propose a method for allowing artists to create parametric svBRDF representations of nearly-flat surfaces using simple printed fiducial markers and video obtained from a mobile phone with a flash that can be constantly illuminated. Our technique does not require any specialized equipment and delivers a representation of the scanned material that is suitable for many artistic applications. We demonstrate the versatility of our method by reproducing a variety of spatially varying materials including leather, fabric, metal, wood, paint, and tile. Figure 1 shows four example materials captured with our method and rendered using Mitsuba [Jak10].

2. Related Work

Generalized surface capture encompasses a wide variety of materials and techniques. Some examples include the Bi-directional Texture Function (BTF) for materials with significant normal variation [DVGNK99], the Bi-directional Scattering Distribution Function (BSDF) for materials with transparency or translucency [BDW81], and the Bi-directional Reflectance Distribution Function (BRDF) for homogeneous and relatively flat opaque materials [Nic65]. In this paper we focus on the spatially varying BRDF (svBRDF), a particular variant of the BRDF also introduced by Nicodemus et al. [NRH*77] that allows for different BRDF parameters at each point on the captured surface.

2.1. Sparse svBRDF Acquisition

A complete sampling of the six-dimensional svBRDF may be performed using a spatial gonioreflectometer [MLH02], although this is a lengthy and data-intensive task. Efforts have been made to simplify this process while still accurately capturing all the necessary variation in surface reflectance. Dong et al. [DWT*10] proposed a hand-held array of LEDs mounted to a camera, along with interpolated svBRDF estimation based on manifold bootstrapping. Aittala et al. [AWL13] used structured light with basis illumination to estimate reflectance. Francken et al. [FCMB09] also obtain high-quality surface normals and gloss estimation from structured LED lighting. Similarly, Ghosh et al. [GCP*09] used structured LEDs with polarizing filters to estimate the reflectance of spherical objects. Zhou et al. [ZCD*16] optimized a sparse blending of sparse basis BRDFs with a limited number of input views. In another minimalist setup, Xu et al. [XNY*16] obtain uniform isotropic BRDFs from the MERL database with a two-shot capture system.

There have also been a variety of capture systems that employ polarized light to separate the diffuse and specular components [MHP*07, GCP*10, RRFG17, TFG*13]. Chen et al. use a linear light source on an electronic rig to estimate anisotropic BRDFs, and Ren et al. [RWS*11] proposed a portable setup involving a static mobile phone, a hand-held linear light source, and a collection of carefully selected materials with known BRDFs. However, all of these methods still require expensive or highly specialized equipment for capture.

2.2. Appearance Matching

Another body of work focuses on tools to help artists match the appearance of a material through user input or by simplifying the material representation. Dong et al. [DTPG11] estimate a simplified model svBRDF for a single texture image and allow users to adjust the behavior of regions of similar appearance until they are satisfied. Di Renzo et al. [DRCP14] produce a layered BRDF plus texture image based on user edits in material space. Xuey et al. [XWT*08] create a static image with material weathering effects for a single lighting environment. Haro et al. [HE03] also produce a static image with a single light source “baked in” to the material appearance. All of these tools circumvent the need for capturing multiple lighting and viewing angles in favor of simplified appearance estimation.

Aittala et al. [AWL15] combined texture synthesis from a no-flash photo with reflectance capture from a single flash photo to produce an svBRDF and normal map, however their technique was limited to highly regular, repeated textures. In subsequent work they replicated these results using a single flash image and deep learning techniques, but with less consistent results [AAL16]. Most recently, Li et al. [LDPT17] also used deep learning to estimate the ambient lighting and thereby generate a diffuse, specular, and normal map decomposition of a single arbitrary image.

2.3. Image-Based Approximation

Our work is most closely related to a group of approaches that approximate a full svBRDF model using a limited set of input images.

Wang et al. [WZT*08] produce anisotropic svBRDFs by synthesizing a spatially varying Normal Distribution Function (NDF) from sparse light positions. Zickler et al. [ZREB06] estimate an isotropic svBRDF with known geometry using a fixed camera with a moving point light source in a controlled lighting environment. Goldman et al. [GCHS10] use a simplified studio setup with multiple high resolution photographs as well as BRDF clustering to estimate the svBRDF of an object. Lensch et al. [LKG*03] also used BRDF clustering, and their iterative subclustering method is similar to ours, although our capture setup is much simpler. Similarly, Zhou et al. [ZWT13] estimate an isotropic svBRDF for an arbitrary-shaped object by employing structure from motion (SFM), a ring light source, and a linear combination of basis BRDFs. Two approaches have been proposed for estimating large scale geometry, specularity, and diffuse albedo based on input from a light field camera [WCER16] and a small number of rendered input images [HS17].

In the space of mobile phone capture, Thanikachalam et al. [TBF*17] estimated reflectance from video and optimized the sampling density and capture path, but with very low resolution output. The approach proposed by Hui and colleagues [HSL*17] requires capturing several images of a texture sample from different viewpoints with the flash providing illumination and then using a dictionary-based approach to select a BRDF that matches the observations of each pixel. They provide an elegant proof showing that it is insufficient to fit a BRDF using only observations where the light is collocated with the camera, as is the case for a cellphone, but by using a dictionary they are able to still obtain plausible results for cases where the subject materials match an example stored in their library. We overcome this limitation in our approach by allowing the user to add a second cellphone camera to obtain observations from other perspectives that are not collocated with the light. This approach allows us to fit a BRDF model directly to the data so that each pixel's appearance is not restricted to an existing dictionary of materials. However, even when only one cell phone is used, our initialization strategy still allows the fitting process to obtain reasonable results.

Finally, Riviere et al. [RPG16], also demonstrated svBRDF capture using mobile phone video. Our proposed method improves on their capture system in two ways. First, our unique video frame alignment technique allows us to capture reflectance data from a much closer distance (10cm vs 50cm) and does not require the entire sample to be visible in each input image. By stitching together many partial observations from a closer view distance, we can obtain very high resolution results even for large samples. We have found that high resolution is generally required to obtain good results when rendering the svBRDFs on objects in 3D scenes, particularly for materials with fine-scale specular features such as gold thread or metallic flakes. The closer viewing distance also produces more oblique lighting and viewing angles and a brighter specular highlight, allowing us to accommodate capture under more varied ambient lighting conditions. We captured most of our data sets under approximately 400 LUX ambient illumination (compared to 40 LUX by Riviere et al.). Second, our method does not require either radiometric calibration of the device nor the inclusion of a specific color chart at capture time. Our fiducial markers are less than 2cm square and can be printed anywhere and easily carried in a wallet. These differences expand the possible use cases of casual svBRDF estimation to more varied lighting environments and more accessible tools for capture.

3. Approximate svBRDFs With Mobile Phone Video

Our proposed capture and fitting technique requires only one or two cell phone cameras with continuous flash video capability and a set of four small fiducial markers which may be printed on standard copy paper. Using these commonly available tools, we are able to fit an svBRDF to mostly-flat, opaque surfaces that have spatially varying reflectance and uneven surface texture.

We first place the fiducial markers around the area of interest and capture a short, hand-held flash video at a relatively fixed distance over the object surface. We then align and warp the resulting video frame images into a single panorama in the global coordinate space with observations from multiple light and view locations for each pixel. In the second step, we cluster the pixels by similar appearance, fit a BRDF to the clusters, and then recursively sub-divide and fit a BRDF and normal vector displacement to each sub-cluster until the residual fit error at each pixel is sufficiently low.

Our output is a high-resolution svBRDF based on the Ward model [War92] that can be easily applied to 3D objects for rendering in applications such as Mitsuba. Additionally, because we do not require the fiducial markers to be visible in each frame, we can capture larger regions at a closer distance than previous work, enabling us to obtain a high-resolution output with a more well-defined specular lobe in each sample image.

3.1. Alignment and Pose Estimation

Aligning the video frames is essentially a panorama stitching problem, but for our application the quality of the alignment must be very precise. Although a traditional panorama need only avoid noticeable seams and distortions, in our case every pixel location needs to be correctly matched across all light and view positions to avoid spurious correlations between lighting and appearance.

The use of a mobile phone camera for this task creates several difficulties that must be overcome for good quality results. Mobile phone camera lenses are usually wide-angle, wide-aperture, and fixed focal length, with a very narrow depth of field (DOF) and significant barrel or moustache distortion. Traditionally, the lens distortion would be corrected using a checkerboard calibration technique [Zha00], but such techniques require either a wide DOF or a relatively large viewing distance so that the entire checkerboard is in focus for all rotated camera positions. Furthermore, in our case it is necessary to use auto-focus to accommodate the hand-held camera motion, but lens "breathing" effects are known to cause lens distortion to vary dramatically across different focus states. Stitching the panorama, therefore, requires solving for both the camera pose and the current lens distortion for every video frame independently.

One possible solution for correcting the distortion would be to use a parametric lens model in conjunction with a homography for the camera pose. However, in practice we found that typical low-order models with two or three parameters were not sufficiently

Alignment Parameters

Scale	SURF	SIFT	Max. Dist.	Min. Inliers
0.25	1,200	3,000 (1,200)	20 pixels	20 points
1	4,800	12,000 (4,800)	60 pixels	60 points

Table 1: The parameters used for the coarse and fine alignment steps. Column 1 shows the scale of the input image (coarse or fine). Columns 2 and 3 show the number of SURF and SIFT features extracted from each image (the number of selected SIFT features for matching is shown in parentheses). Columns 4 and 5 show the maximum distance between matched inlier points and the minimum number of inlier points for the MLESAC procedure, respectively.

accurate for our application, and higher-order models with up to seven parameters were cumbersome, slow to converge, and easily derailed by local minima. Another solution is to include fiducial markers to establish known physical locations from which one might compute a homography for each frame. To undistort the entire image, the markers would need to be placed on the outside edges of the captured area and all the markers would need to be visible in each video frame. Even for a relatively small sample of 10x10 cm, this arrangement requires a capture distance of at least 30 cm. However, we found that a capture distance closer to 10-15 cm was more ideal because it produces more oblique lighting and viewing angles, provides a brighter specular highlight, and allows for a higher resolution image of the surface for BRDF fitting.

In order to capture larger areas at a close distance, we instead perform a feature-based alignment for each frame using multiple overlapping homographies to obtain a piece-wise linear approximation of the lens distortion, as explained in section 3.1.4. We establish a global coordinate system by placing four printed fiducial markers at the far edges of the captured region and obtaining a single reference image of the entire area. The homography solution for each sub-frame is then calculated relative to this global space, allowing us to estimate the 3D world coordinate camera and light positions for all frames, even though 50-80% of frames have no fiducial markers visible.

3.1.1. Fiducial Markers and Reference Image

Fiducial markers were created and detected using the ArUco "6x6_250" predefined library [GJMSMCMJ14]. The actual size of the printed markers was 1.6 cm square. Four markers were placed on the flat capture surface, one in each corner of the area to be captured. An additional reference image containing all four markers was also captured at a distance of approximately 20 cm perpendicular to the surface. The android app "Level Camera" was used to ensure the camera sensor was parallel to the surface for the reference image [Wen13]. The locations of the fiducial marker corner points were recorded separately for both the reference image and in each video frame where the fiducials were visible.

3.1.2. Removing Blurry and Disconnected Frames

Because the camera motion is hand-held without any physical guide or apparatus, irregular motion can sometimes produce blurry frames as a result of intermittent defocus or motion blur. To detect blurry frames we computed a blur metric f based on the power spectrum for each image i such that

$$f(i) = \text{mean}(\log_{10}(0.001 + |\mathcal{F}(i)|) \qquad (1)$$

where $\mathcal{F}(i)$ is the Fourier transform of the image, and the absolute value, log, and addition operators are applied competent-wise and the mean is taken across the result. Frames with a value of $f(i)$ less than 1.5σ below the mean across all frames were discarded.

When removing frames due to blur or insufficient feature matches, there is a potential for a small subset of frames to be well-connected to each other but lack at least four points of connection to any other frame in the sequence. In that case it is impossible to determine an unambiguous projective transformation of that subset to the global space. At the end of the feature matching process we therefore obtain the connected sub-graph of the connectivity map with the most members and remove any frames not contained in that sub-graph.

3.1.3. Coarse Alignment

Although each video frame may be trivially assumed to overlap with its immediate neighbors in time, accurate stitching of a full panorama also requires accurate loop closure for non-neighboring frames. However, feature matching across all pairs of frames at full resolution is costly and also likely to return many false matches for self-similar textures. We therefore first perform a coarse alignment step at a subsampled scale to determine overlapping frames, then repeat the process for the full resolution images to obtain the locations of matching features for the final homography estimation. Parameters for both alignment steps are shown in Table 1.

For the coarse alignment step, each frame was downsampled 4x, and a maximum of up to 1,200 uniformly distributed SURF features [BETVG08] and 3,000 SIFT features [Low04] were extracted from each frame. SIFT features were obtained and matched using the CudaSIFT library [BBK14]. Features within a 75 pixel radius of the center of the image were discarded to avoid false matches of the specular highlight. During the feature matching process, all the SURF features and a random subset of 1200 SIFT features were uniquely matched (1 to 1) to all the features from each other frame. The matched feature points were used to estimate a similarity transformation between each pair of frames using MLESAC [TZ00], with a maximum distance of 20 pixels between inlier feature locations. Any number of inliers greater than 20 was recorded as a potential match.

The resulting matrix of inlier counts (the connectivity map) was further thresholded and filtered to remove spurious matches. The threshold for the minimum number of inliers was determined by the 50th percentile of those frame pairs with some overlap. This ensured that no more than 50% of all frames could be overlapping and only the strongest connections remained. Finally, the connectivity map was smoothed using a 5x5 median filter to remove any non-continuous matches.

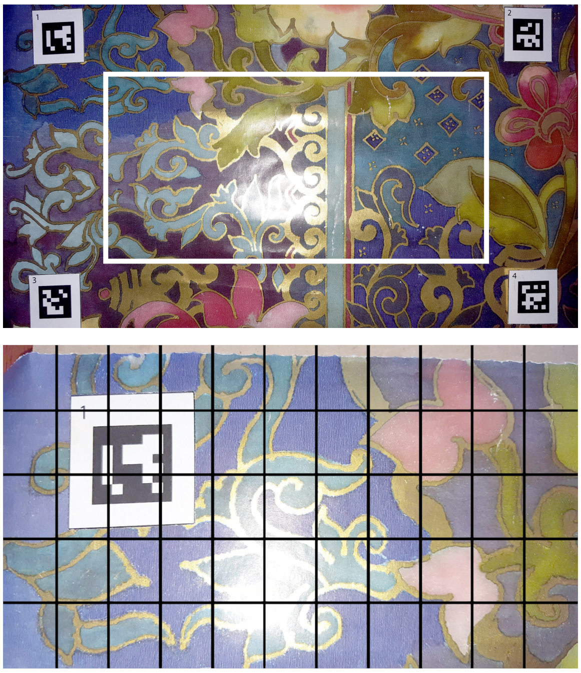


Figure 2: The top subfigure shows an example reference image with the user-selected output region outlined in white, while the bottom subfigure shows an example video frame from the same data set with sub-frame boundaries overlaid. The contrast of the example video frame has been decreased for visualization. The width of the black lines indicates the overlap between adjacent sub-frames.

3.1.4. Fine alignment and subdividing frames

In the fine alignment step, full-scale feature point locations are divided into sub-frame regions and we obtain a global least-squares solution for the homography transformation of each sub-frame.

Feature matching was only performed for overlapping image pairs from the coarse alignment connectivity map, with slightly modified parameters. The flash feature removal radius, maximum number of SURF features, and the max number of SIFT features were all scaled up by 4x. The maximum feature location distance for MLESAC was set at 60 pixels, and the minimum number of inliers was 60. The large allowable MLESAC distance error is a reflection of the amount of lens distortion. Although larger allowable error may cause incorrect matching, restricting the inliers to only precise matches causes only the undistorted portions of each frame to be matched, and this defeats the purpose of the alignment process completely. It is therefore much better to have a larger distance error and enforce precision by increasing the number of inliers. Ideally, any remaining false matches are greatly outnumbered in the final least squares minimization process.

The inliers from each frame were divided up into a 5x11 grid of uniformly sized sub-frames whose dimensions were determined by empirically examining the level of lens distortion in the phone cameras we used. An illustration of the sub-frame divisions is shown in Figure 2. The size of each sub-frame was 448x370 pixels with an X and Y overlap of 25 and 22 pixels, respectively. Due to similarity of camera hardware specifications across mobile phones, it is likely that these values would be appropriate for other devices as well.

3.1.5. Linear approximation solution

Once we have obtained the corresponding feature locations, we solve for a homography transformation matrix for each sub-frame to the global space defined by the location of the fiducial markers in the reference image.

To obtain the transformation matrices, we perform a global least-squares fit simultaneously for all corresponding feature pairs across all overlapping frames. Our solution is the set of homography matrices that minimizes the sum of squared differences between the projected global positions of each shared feature point p such that

$$\min \sum \|F_{p_i} \cdot H_i - F_{p_j} \cdot H_j\|^2 \tag{2}$$

where F_{p_i} and F_{p_j} correspond to the $[x, y, w]$ homogeneous coordinates of feature point p in each pair of overlapping sub-frames i, j, and H_i and H_j are the corresponding homography matrices that project each image into the global space.

Unraveling and concatenating all homography matrices H_i into a single vector h allows us to construct a large sparse matrix F_{pij} where each column corresponds to one entry of h, and each row corresponds to $p_i - p_j$ in homogeneous coordinates. Our minimization problem then becomes

$$F_{pij} \cdot h = 0. \tag{3}$$

Furthermore, since a homography is only precise up to a scale factor, we add the following constraints to define the global space:

$$H_i(3,3) = 1 \tag{4}$$

such that the (3,3) entry of each homography matrix is defined to be one, and

$$F_m \cdot h_m + F_{\notin m} \cdot h_{\notin m} = 0 \tag{5}$$

$$F_{\notin m} \cdot h_{\notin m} = -k_m \tag{6}$$

where F_m is the set of rows in F_{pij} containing the m fiducial marker points, h_m is the corresponding entries of h, and $F_{\notin m}$ and $h_{\notin m}$ are the remaining entries of F_{pij} and h, respectively. The i entries of F_m are from the marker point locations in each sub-frame, while the j entries are from the marker point locations in the reference image. In (6) the product of the known entries is moved to the righthand side of the equation, yielding $-k_m$, so that $h_{\notin m}$ may be obtained via least squares.

3.1.6. Pose estimation

We determine the real world position of the camera using the homography of the center sub-frame of each input image. Each transformation matrix is decomposed into its component rotation matrix, R_i, and translation vector, t_i according to the process described in Malis et al. [MV07]. We use these components to construct a set of 3D homogeneous matrices as shown in (7), wherein each matrix transforms from the reference image pose to the corresponding camera pose for each frame.

$$\begin{bmatrix} & R_i & & t_i \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{7}$$

The reference image camera pose and light position are determined as follows. The field of view (FOV) of the camera is calculated offline using a one-time calibration image with an object of known size at known distance. Both the rigid offset of the flash relative to the camera and the size of the fiducial markers are also measured offline. In our case FOV was measured to be 70°, the XYZ offset of the flash was [1.4, -0.3, 0] centimeters, and the fiducial markers were each 1.6 centimeters tall.

At capture time, the global XYZ origin point is defined to be the point on the captured surface corresponding to the center of the reference image. The reference camera pose is therefore located at [0, 0] in XY. The Z distance is triangulated from the FOV and the average size of the reference fiducial markers in pixels relative to their known physical size in centimeters. The reference light position is obtained by applying the known flash offset to the reference camera pose.

Finally, the world-to-image transformation matrix described in (7) is applied to both the reference camera and light positions to obtain the camera and light positions for each frame. Unreliable camera poses located outside the fiducial marker boundaries are discarded, along with their corresponding video frames.

3.2. Clustering and BRDF Fitting

For each point on the surface, our sparse input samples typically cover only a very tiny sliver of the 4-dimensional hemisphere used to fit a BRDF. However, most materials are made up of many regions that share similar reflectance properties and would all be well-described by a single BRDF with minimal loss of accuracy. We take advantage of this self-similarity by clustering similar pixels together and fitting a BRDF to each cluster.

Determining the number and size of the clusters presents a tradeoff between generalizability and fidelity to the observed data. There are many ambiguous BRDF solutions that can produce the same appearance behavior. Larger clusters are likely to include a more complete sampling of the BRDF hemisphere and therefore converge to a more accurate representation, but they are also more likely to obscure the small details and variation which make spatially varying materials interesting. If the clusters are too small, however, it is probable that over-fitting will produce an incorrect result which does not generalize to novel light and view positions which were absent from the captured video.

Similar to Lensch et al. [LKG*03], our solution is to initialize the BRDF with very large clusters and a constrained BRDF model, and then recursively subdivide the clusters, initializing each sub-cluster with the fitting output of its parent. Our initial clusters are grouped based on the average observed color of each pixel and then each cluster and sub-cluster is subdivided based on the per-pixel residual of the fitted BRDF. This encourages each smaller sub-cluster to find a solution in the neighborhood of solutions defined by the larger parent cluster, greatly reducing the likelihood of obtaining an incorrect ambiguous solution.

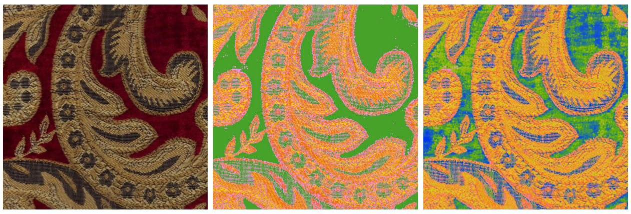


Figure 3: Initial and final sub-clusters for the two-camera red velvet lamé material. The left image shows the average color. Five clusters were obtained in the initial clustering (middle), and the final result included 5,152 sub-clusters (right).

For each sub-cluster we produce an anisotropic Ward svBRDF [War92] and a normal map. We are therefore able to fit opaque materials that do not have any Fresnel effects. Due to our feature-based homography alignment process, we also require the scanned material to be relatively flat and have at least some medium-scale albedo variation to align the input video frames.

3.2.1. Clustering and svBRDF Initialization

Using the aligned images, we coarsely approximate the diffuse albedo color by the average color of each pixel in the global coordinate space. This average color image is then converted to CIE 1976 L*a*b* color space. We then apply k-means clustering with k-means++ initial centroid positions [AV07] to the normalized albedo color values. The number of clusters, k, is chosen based on the linear bisection point of the summed squared euclidean error across all values of k in the range k = [2:20]. For our data, typically k=4.

For each initial cluster, we fit an isotropic BRDF (see section 3.2.3) with a single normal vector for the cluster, constrained to be perpendicular to the surface (that is, $n = [0,0,1]$). This step initializes the ρ_d, ρ_s, and α to reasonable values for the average normal vector orientation. The initial conditions for the isotropic fitting step are the average color over the entire cluster for the the diffuse component (ρ_d) and twice the average of the standard deviation across frames for the specular component (ρ_s). The roughness parameter (α) is initialized to 0.1.

Once an isotropic BRDF has been fit to each initial cluster, we calculate the least squares fit error for each pixel in the cluster and recursively subdivide the pixels into two sub-clusters according to the threshold

$$t = \text{median}(E_{\text{px}}) + \text{mad}(E_{\text{px}}) \tag{8}$$

where mad is the median absolute deviation and E_{px} is the per-pixel fit error averaged over all observations for each pixel. Each sub-cluster is then fit with a full anisotropic BRDF and a normal offset, and the per-pixel fit error is calculated for the next iteration. We continue to subdivide clusters in this way until we reach a minimum cluster size of 50 pixels. Figure 3 shows an example of the progression from initial to final clusters for the red velvet lamé material.

3.2.2. Redundant Observations

Larger clusters tend to contain many redundant observations of similar materials from almost identical viewpoints. These extra observations dramatically increase the BRDF optimization runtime without improving the accuracy of the fit. To simplify our fitting process, we apply a binning and averaging step to obtain unique viewpoints. At each sub-clustering iteration, we group all observations for all pixels in the subcluster into 5° increments for each of θ_i, ϕ_i, θ_r, and ϕ_r, and 1 cm increments for the light radius, r. For each unique combination of these variables, all the BRDF input parameters (including light and view positions and observed color) are averaged together into a single unique observation for fitting. The contribution of each unique observation to the calculated fitting error is then weighted by the number of raw observations included in its average, according to Equation 11. To calculate the per-pixel fitting error, the fitted value for each unique viewpoint is applied to all the raw observations in its group.

3.2.3. Reflectance Modeling

We model the surface appearance for each point as the incident light from the camera flash multiplied by the BRDF and modulated by the solid angle of the light source as a function of the incident light angle.

The surface appearance is therefore described as

$$L_r(\theta_r,\phi_r) = \int_0^{2\pi}\int_0^{\frac{\pi}{2}} L_i\cdot(\theta_i,\phi_i)\cdot\rho_{bd}(\theta_i,\phi_i;\theta_r,\phi_r)\cdot\frac{\cos(\theta_i)\cdot dA}{r^2} \tag{9}$$

where

L_r is the reflected radiance to the camera
L_i is the incident radiance from the light
ρ_{bd} is the BRDF
θ_r and ϕ_r are the camera elevation and azimuth angles
θ_i and ϕ_i are the light elevation and azimuth angles
dA is the differential surface area of each pixel
r^2 is the radial distance to the light source

and all angles are relative to the normal vector. Similar to Aittala et al. [AWL15] and many others, the ambient light is not explicitly modeled but rather implicitly incorporated into the BRDF.

The ρ_{bd} term in (9) is the Ward BRDF model, described by the following equation

$$\rho_{bd}(\theta_i,\phi_i;\theta_r,\phi_r) = \frac{\rho_d}{\pi} + \frac{\rho_s\cdot e^{-\tan^2(\theta_h)\cdot\left(\frac{\cos^2(\phi_h)}{\alpha_x^2}+\frac{\sin^2(\phi_h)}{\alpha_y^2}\right)}}{4\pi\cdot\alpha_x\cdot\alpha_y\cdot\sqrt{\cos(\theta_i)\cdot\cos(\theta_r)}} \tag{10}$$

where

ρ_d and ρ_s are the diffuse and specular albedo values
α_x and α_y are the roughness parameters in X and Y
θ_h and ϕ_h are the elevation and azimuthal angles of the half-vector between the light and camera

In the initial clustering step, an isotropic variant of this model is used wherein $\alpha_x = \alpha_y$. Subsequent subclustering iterations are fitted using the full anisotropic BRDF model and two normal vector offset angles, n_{θ_x} and n_{θ_z}, which describe the rotation of the normal vector about the X and Z axes respectively. In the final svBRDF and normal map, all the pixels in each sub-cluster are therefore represented by the eight BRDF parameters above (one per color channel for ρ_d and ρ_s) and two normal vector offset parameters.

Our optimization problem is therefore

$$\begin{array}{lll} \text{minimize} & \sum w\cdot\sum(L_f-L_o)^2 & \\ \text{subject to} & \{\rho_d,\rho_s\}\geq 0 & 0^\circ \leq n_{\theta_x} \leq 45^\circ \\ & \{\alpha_x,\alpha_y\} > 0 & 0^\circ \leq n_{\theta_z} \leq 180^\circ \\ & \rho_d+\rho_s \leq 1 & \end{array} \tag{11}$$

where L_o is the observed color values, L_f is the fitted BRDF evaluated at corresponding angles to L_o, and w is the number of samples per unique viewpoint as described in section 3.2.2. We solve for (11) using a sequential quadratic programming (SQP) optimization function [NW06].

3.3. Joining Two Video Streams

Although the reflectance properties of many materials are well-described by observations using a single collocated camera and light source, incorporating a second simultaneous video stream allows us to also capture somewhat more complex materials without requiring other specialized tools. By capturing one video with the camera flash turned on alongside a second no-flash video, we can observe the behavior of the scanned material at more oblique light and view angles and thereby obtain a more complete sampling of the BRDF.

The majority of our pipeline is image-based and accepts a second video stream without any modification. Our only requirement is that the two video streams be temporally synchronized at the first frame of each video, and that the length of the no-flash video be shorter than the flash video. This ensures that the position of the light source is known for all observed input frames.

To synchronize the time streams, we simply begin the no-flash recording first and then crop the start of the no-flash video to the frame where the light from the flash camera first appears. At the frame rates used in our capture setup the actual transition frame is typically highly visible because the rolling shutter effect produces an obvious transition line across the frame. This method afforded acceptable synchronization for our application where the hand held cameras are moving relatively slowly.

4. Results

Our capture data were obtained using a Samsung Galaxy S6 (or S7 for the second camera). The resolution of the reference images was 5312x2988, and the videos were captured at a rate of 30 frames per second (fps) and resolution of 2160x3840 pixels. We captured video in Pro-Mode with the flash turned on, using a shutter speed of 1/500 seconds and an appropriate ISO setting for the ambient light level, between 50 and 400. White balance was manually set at the beginning of each video to ensure consistency across frames.

The camera was moved over the surface by hand in a sinusoidal top-to-bottom and side-to-side fashion to achieve relatively even coverage of the entire captured area. Typical video capture distance was between 10-20 centimeters from the surface, and reference image capture distance was usually around 20-30 centimeters.

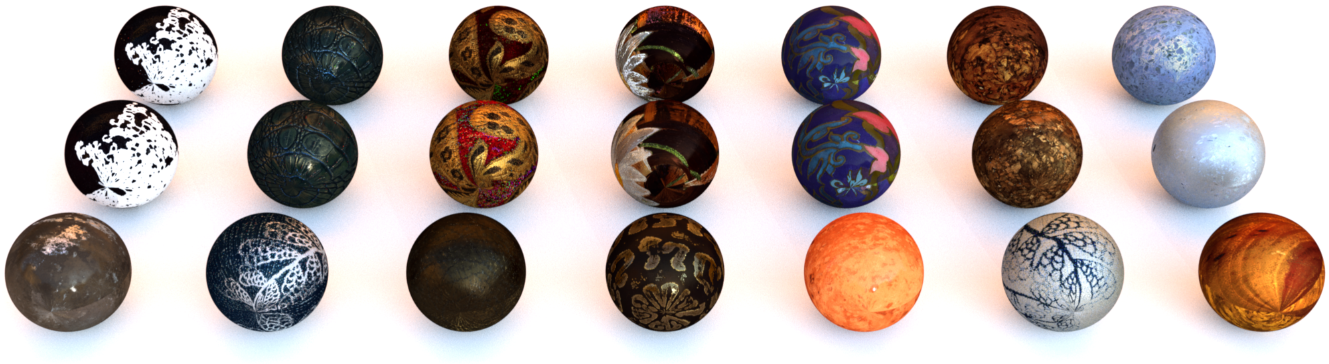


Figure 4: All of the scanned materials rendered onto spheres with Mitsuba and illuminated by the Pisa environment map and a single additional point light source. The top row is captured with two cameras, the middle row depicts the same materials captured with a single camera, and the bottom row shows additional materials captured with only one camera. From top to bottom and left to right: abstract oil painting, green faux leather, red velvet lamé, woven rattan mat, wrapping paper, corkboard, shiny white tile, aged metal patina, blue damask fabric, buffalo leather, metallic embossed paper, orange ceramic tile, damask fabric reversed, and wood block.

The supplemental material includes a diagram of our capture setup. Each video was 15-30 seconds in duration, covering an area of approximately 20x20 cm. From this sequence we extracted every 5th frame (6 fps) for the single-camera examples and every 10th frame (3 fps) for the two-camera examples. We found this sampling rate to be an acceptable trade-off between sampling density and data processing limitations.

We provide examples of seven materials captured with only a single camera (aged metal patina, blue damask fabric, buffalo leather, metallic embossed paper, orange ceramic tile, damask fabric reversed, and wood block), and seven materials captured with both one and two cameras for comparison (abstract oil painting, green faux leather, red velvet lamé, woven rattan mat, wrapping paper, corkboard, and shiny white tile).

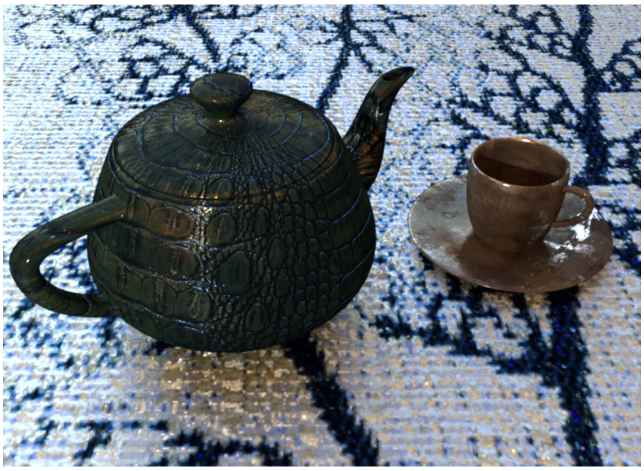


Figure 5: A natural scene with three scanned materials rendered with Mitsuba and illuminated by the Pisa environment map and a single additional point light source. The table surface is textured with the damask fabric reversed material, the teapot is textured with the faux green leather material, and the teacup is textured with the aged metal patina material.

Figure 4 shows a rendering of all the captured materials mapped onto spheres and illuminated by the Pisa environment map with a single additional point light source. The examples in the top row are captured with two cameras, the middle row depicts the same materials captured with a single camera, and the bottom row shows additional materials captured with only one camera. Figure 5 also shows several of the same materials used in a more natural scene under the the same illumination.

In Figure 6 we include a comparison to ground truth for the example materials from Figure 1 using a very oblique light position (the supplemental material includes the same visualization for all other materials). This comparison is very challenging because the lighting configuration is very different from anything in the input data for fitting the svBRDF, so our algorithm must rely on smoothness assumptions implicit in the Ward model. It is apparent that some high frequency texture and corresponding specular highlights are missing for several materials. These highlights most likely occupy a very sharp peak of the BRDF, and are thus difficult for any method to accurately reproduce without direct observation. Nonetheless our method produces a plausible appearance for these samples. Additionally, in a supplemental video we also show a comparison between the input video frames and a rendering of the fitted svBRDF output using the input light and camera locations, for all materials.

Each of the svBRDF output layers is also included for more detailed analysis. Figure 7 shows the raw results for the same subset of example materials. The leftmost column is the average color as described in section 3.2.1. The remaining columns are the diffuse color (ρ_d), the specular color (ρ_s), the roughness parameter in the X direction (α_x), the roughness parameter in the Y direction (α_y), and the normal offset map. For materials captured with both one and two cameras, the results are shown side by side for comparison.

The supplemental material also includes a second video showing the materials textured onto a 3D object with animated illumination changes, as well as a comparison between the one- and two-camera results.

The differences between the quality of the single and dual camera results for the red velvet lamé and wrapping paper materials reveal the importance of broader sampling for more complex materials. The diffuse parameter color is slightly darker for the two-camera wrapping paper example, but the overall result is very similar to the one camera result. However, for the red velvet lamé, the single camera case has much more trouble separating and distinguishing reflectance behavior that changes quickly with direction, as predicted by Hui et al. [HSL*17]. We still get usable results with a single camera, but the algorithm is unable to disambiguate between a bright surface tilted away from the camera and a darker surface tilted toward the camera, resulting in over-fitting to the data. This problem could potentially be corrected manually, but given that it is relatively easy to use two cellphones, we feel that two cameras is the preferred option when accurate reproduction is desired.

5. Conclusion and Future Work

We have demonstrated a new technique for capturing and modeling the appearance of nearly flat surfaces using printed fiducial markers and a mobile phone video with continuous flash. Our technique employs a very simple capture process using off-the-shelf hardware, and the output of our system may be directly textured onto 3D objects using standard rendering software. We have also provided examples showing a variety of materials captured with both one and two cameras and rendered under complex lighting environments.

Our technique has several limitations. Because we align the video frame images using homographies, we are only able to capture flat surfaces with relatively minimal surface relief. Our feature-based alignment also requires captured materials to have some irregular, medium-scale textural variation, which means we are unable to align extremely repetitive textures. However, Aittala and colleagues have already proposed two excellent solutions for capturing self-similar materials and we consider our work to be complementary to theirs [AWL15, AAL16].

The results shown in this paper are generated with very sparse sampling and a very simple BRDF model, and we are therefore unable to capture phenomena such as Fresnel effects. However, we place no restrictions on the model used, and it has been suggested that some micro-facet models may be better suited to approximating more complex reflectance behavior [BLPW14]. Our panorama stitching method could also be combined with the dictionary approach proposed by Hui and colleagues [HSL*17] to obtain high-resolution models of complex materials that require sampling at very oblique light and camera angles. For materials with much more complicated reflectance, our implementation would easily allow the second no-flash camera to be placed on a tripod at an oblique angle to the surface to capture the entire flash sequence from the side.

Finally, our research-quality code is not yet optimized and takes about 2 hours to align and fit an svBRDF from a 20 second video, half of which is taken up by feature extraction and matching. We speculate that using optic flow information for loop closure might produce a better estimation of overlap across frames without the need for feature matching in the coarse alignment step, providing a significant time savings.

However, on the whole we believe that the reduced equipment requirements and simplicity of our capture methodology are a valuable contribution to the state of the art. Any content creator with access to a printer and a mobile phone can quickly and easily capture a variety of interesting materials encountered in everyday life. We hope that our system will inspire more widespread capture of real-world materials and encourage future development of svBRDF fitting techniques using sparse data.

References

[AAL16] Aittala M., Aila T., Lehtinen J.: Reflectance modeling by neural texture synthesis. *ACM Transactions on Graphics (TOG) 35*, 4 (2016), 65. 2, 9

[AV07] Arthur D., Vassilvitskii S.: k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms* (2007), Society for Industrial and Applied Mathematics, pp. 1027–1035. 6

[AWL13] Aittala M., Weyrich T., Lehtinen J.: Practical SVBRDF capture in the frequency domain. *ACM Trans. Graph. 32*, 4 (2013), 110. 2

[AWL15] Aittala M., Weyrich T., Lehtinen J.: Two-shot SVBRDF capture for stationary materials. *ACM Transactions on Graphics (TOG) 34*, 4 (Aug. 2015), 110. 2, 7, 9

[BBK14] Björkman M., Bergström N., Kragic D.: Detecting, segmenting and tracking unknown objects using multi-label mrf inference. *Computer Vision and Image Understanding 118* (2014), 111–127. 4

[BDW81] Bartell F. O., Dereniak E. L., Wolfe W. L.: The Theory And Measurement Of Bidirectional Reflectance Distribution Function (Brdf) And Bidirectional Transmittance Distribution Function (BTDF). vol. 0257, pp. 154–160. 2

[BETVG08] Bay H., Ess A., Tuytelaars T., Van Gool L.: Speeded-up robust features (SURF). *Computer vision and image understanding 110*, 3 (2008), 346–359. 4

[Ble05] Blevins N.: Leather Material, June 2005. 1

[Ble13] Blevins N.: Layering Materials, Sept. 2013. 1

[BLPW14] Brady A., Lawrence J., Peers P., Weimer W.: genbrdf: discovering new analytic brdfs with genetic programming. *ACM Trans. Graph. 33* (2014), 114:1–114:11. 9

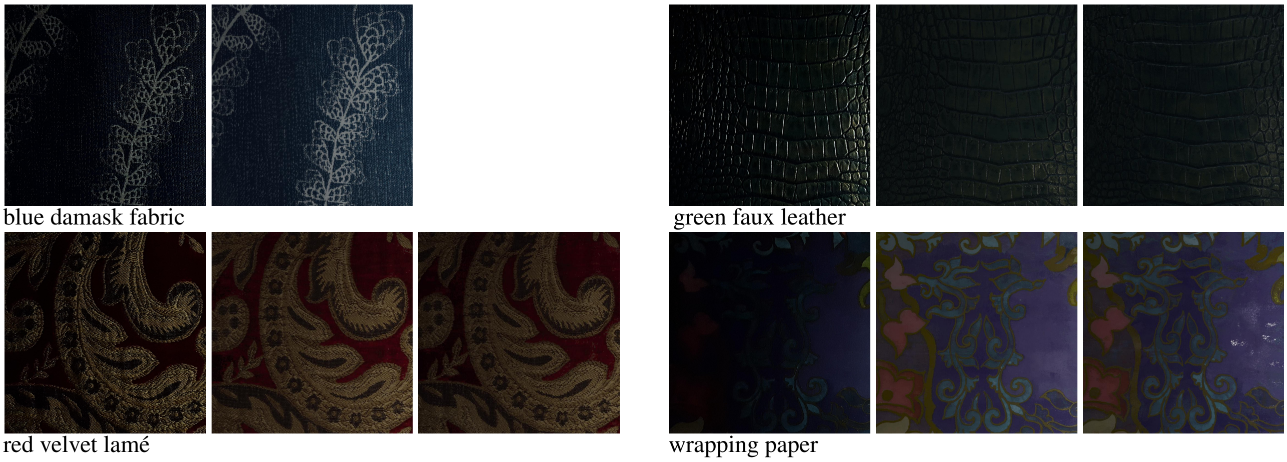
[DRCP14] Di Renzo F., Calabrese C., Pellacini F.: AppIm: linear spaces for image-based appearance editing. *ACM Transactions on Graphics (TOG) 33*, 6 (2014), 194. 2

[DTPG11] Dong Y., Tong X., Pellacini F., Guo B.: AppGen: interactive material modeling from a single image. In *ACM Transactions on Graphics (TOG)* (2011), vol. 30, ACM, p. 146. 2

[DVGNK99] Dana K. J., Van Ginneken B., Nayar S. K., Koenderink J. J.: Reflectance and texture of real-world surfaces. *ACM Transactions on Graphics (TOG) 18*, 1 (1999), 1–34. 2

[DWT*10] Dong Y., Wang J., Tong X., Snyder J., Lan Y., Ben-Ezra M., Guo B.: Manifold bootstrapping for SVBRDF capture. In *ACM Transactions on Graphics (TOG)* (2010), vol. 29, ACM, p. 98. 2

[FCMB09] Francken Y., Cuypers T., Mertens T., Bekaert P.: Gloss and normal map acquisition of mesostructures using gray codes. In *International Symposium on Visual Computing* (2009), Springer, pp. 788–798. 2


blue damask fabric

green faux leather

red velvet lamé

wrapping paper

Figure 6: Comparison to a ground truth photo with an oblique light angle not included in the input fitting data. For each material shown, the first image is the ground truth and the second image is a rendering with the same light pose as the ground truth using the data captured with one camera, and the third image (if available) shows the same rendering using the data captured with two cameras. Images have been cropped square and resized to fit.

[GCHS10] Goldman D. B., Curless B., Hertzmann A., Seitz S. M.: Shape and Spatially-Varying BRDFs from Photometric Stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence 32*, 6 (June 2010), 1060–1071. 3

[GCP*09] Ghosh A., Chen T., Peers P., Wilson C. A., Debevec P.: Estimating specular roughness from polarized second order spherical gradient illumination. In *SIGGRAPH 2009: Talks* (2009), ACM, p. 30. 2

[GCP*10] Ghosh A., Chen T., Peers P., Wilson C. A., Debevec P.: Circularly polarized spherical illumination reflectometry. In *ACM Transactions on Graphics (TOG)* (2010), vol. 29, ACM, p. 162. 2

[GJMSMCMJ14] Garrido-Jurado S., MuÃśoz-Salinas R., Madrid-Cuevas F. J., MarÃ■n-JimÃľnez M. J.: Automatic generation and detection of highly reliable fiducial markers under occlusion. *Pattern Recognition 47*, 6 (2014), 2280–2292. 4

[HE03] Haro A., Essa I. A.: Exemplar-Based Surface Texture. pp. 95–101. 2

[HS17] Hui Z., Sankaranarayanan A. C.: Shape and Spatially-Varying Reflectance Estimation from Virtual Exemplars. *IEEE Transactions on Pattern Analysis and Machine Intelligence 39*, 10 (Oct. 2017), 2060–2073. 3

[HSL*17] Hui Z., Sunkavalli K., Lee J.-Y., Hadap S., Wang J., Sankaranarayanan A. C.: Reflectance Capture Using Univariate Sampling of BRDFs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 5362–5370. 3, 9

[Jak10] Jakob W.: Mitsuba renderer, 2010. http://www.mitsuba-renderer.org. 2

[LDPT17] Li X., Dong Y., Peers P., Tong X.: Modeling Surface Appearance from a Single Photograph Using Self-augmented Convolutional Neural Networks. *ACM Trans. Graph. 36*, 4 (July 2017), 45:1–45:11. 2

[LKG*03] Lensch H., Kautz J., Goesele M., Heidrich W., Seidel H.-P.: Image-based reconstruction of spatial appearance and geometric detail. *ACM Transactions on Graphics (TOG) 22*, 2 (2003), 234–257. 3, 6

[Low04] Lowe D. G.: Distinctive image features from scale-invariant keypoints. *International journal of computer vision 60*, 2 (2004), 91–110. 4

[Mat03] Matusik W.: *A data-driven reflectance model*. PhD thesis, Massachusetts Institute of Technology, 2003. 1

[MHP*07] Ma W.-C., Hawkins T., Peers P., Chabert C.-F., Weiss M., Debevec P.: Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination. In *Proceedings of the 18th Eurographics conference on Rendering Techniques* (2007), Eurographics Association, pp. 183–194. 2

[MLH02] McAllister D. K., Lastra A., Heidrich W.: Efficient rendering of spatial bi-directional reflectance distribution functions. In *Proceedings of the ACM SIGGRAPH/EUROGRAPHICS conference on Graphics hardware* (2002), Eurographics Association, pp. 79–88. 2

[MV07] Malis E., Vargas M.: *Deeper understanding of the homography decomposition for vision-based control*. PhD Thesis, INRIA, 2007. 5

[Nic65] Nicodemus F. E.: Directional Reflectance and Emissivity of an Opaque Surface. *Applied Optics 4*, 7 (July 1965), 767. 2

[NRH*77] Nicodemus F. E., Richmond J. C., Hsia J. J., Ginsberg I. W., Limperis T.: *Geometrical considerations and nomenclature for reflectance*, vol. 160. US Department of Commerce, National Bureau of Standards Washington, DC, USA, 1977. 1, 2

[NW06] Nocedal J., Wright S.: *Numerical optimization*. Springer Science & Business Media, 2006. 7

[Pri16] Price A.: Introducing Poliigon - our new texture site!, May 2016. 1

[RPG16] Riviere J., Peers P., Ghosh A.: Mobile surface reflectometry. In *Computer Graphics Forum* (2016), vol. 35, Wiley Online Library, pp. 191–202. 3

[RRFG17] Riviere J., Reshetouski I., Filipi L., Ghosh A.: Polarization imaging reflectometry in the wild. In *ACM Transactions on Graphics (TOG)* (Nov. 2017), vol. 36, ACM, p. 206. 2

[RWS*11] Ren P., Wang J., Snyder J., Tong X., Guo B.: Pocket reflectometry. *ACM Transactions on Graphics (TOG) 30*, 4 (Aug. 2011), 45. 2

[TBF*17] Thanikachalam N., Baboulaz L., Firmenich D., SÃijsstrunk S., Vetterli M.: Handheld reflectance acquisition of paintings. *IEEE Transactions on Computational Imaging PP*, 99 (2017), 1–1. 3

[TFG*13] Tunwattanapong B., Fyffe G., Graham P., Busch J., Yu X., Ghosh A., Debevec P.: Acquiring reflectance and shape from continuous spherical harmonic illumination. *ACM Transactions on graphics (TOG) 32*, 4 (2013), 109. 2

[TZ00] TORR P. H., ZISSERMAN A.: MLESAC: A new robust estimator with application to estimating image geometry. *Computer Vision and Image Understanding 78*, 1 (2000), 138–156. 4

[War92] WARD G. J.: Measuring and modeling anisotropic reflection. *ACM SIGGRAPH Computer Graphics 26*, 2 (1992), 265–272. 3, 6

[WCER16] WANG T.-C., CHANDRAKER M., EFROS A. A., RAMAMOORTHI R.: SVBRDF-Invariant Shape and Reflectance Estimation From Light-Field Cameras. pp. 5451–5459. 3

[Wen13] WENZL S.: Level Camera - Picture Series - Crunchy ByteBox, 2013. 4

[WZT*08] WANG J., ZHAO S., TONG X., SNYDER J., GUO B.: Modeling anisotropic surface reflectance with example-based microfacet synthesis. In *ACM Transactions on Graphics (TOG)* (2008), vol. 27, ACM, p. 41. 3

[XNY*16] XU Z., NIELSEN J. B., YU J., JENSEN H. W., RAMAMOORTHI R.: Minimal BRDF Sampling for Two-shot Near-field Reflectance Acquisition. *ACM Trans. Graph. 35*, 6 (Nov. 2016), 188:1–188:12. 2

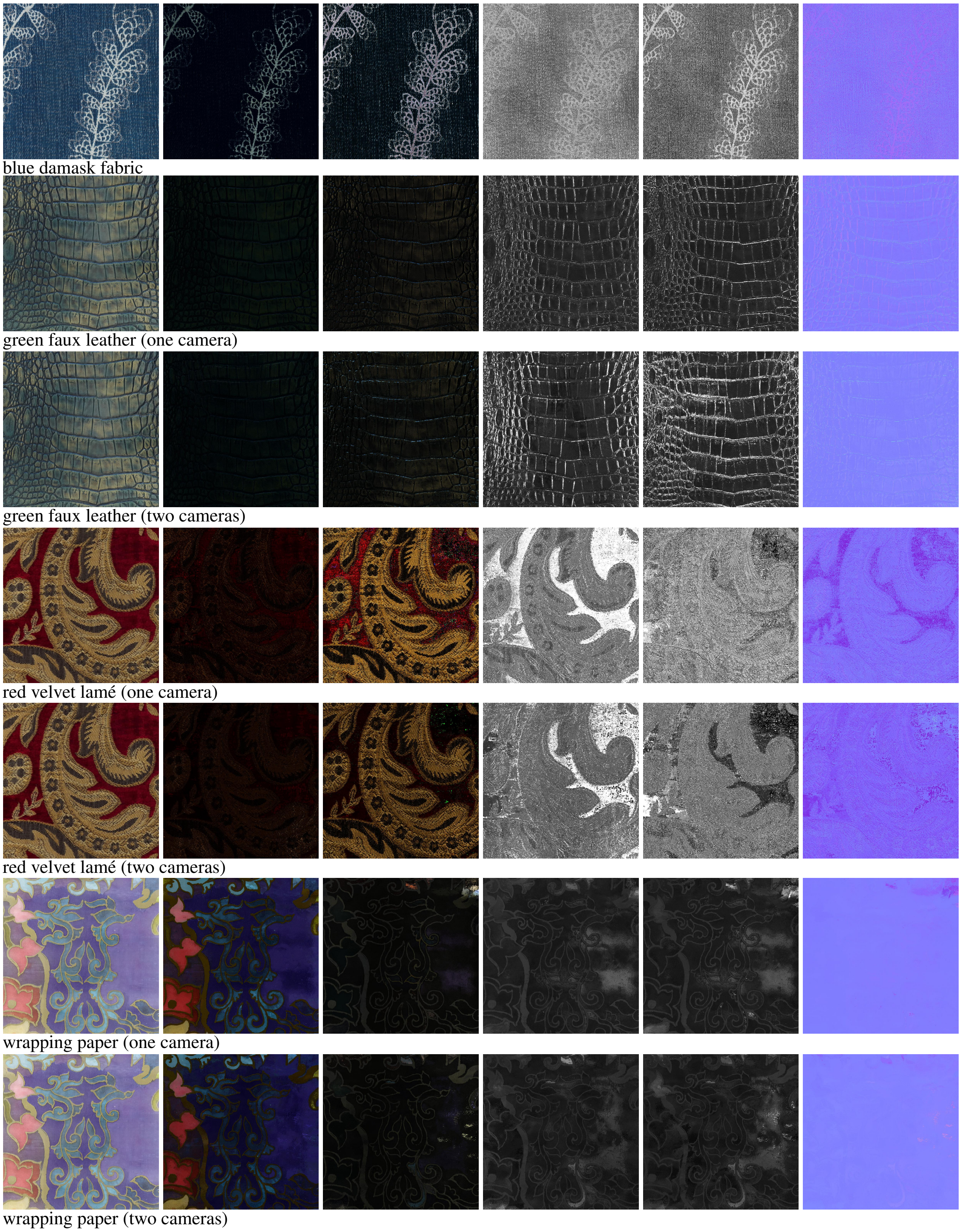
[XWT*08] XUEY S., WANG J., TONG X., DAI Q., GUO B.: Image-based Material Weathering. In *Computer Graphics Forum* (2008), vol. 27, Wiley Online Library, pp. 617–626. 2

[ZCD*16] ZHOU Z., CHEN G., DONG Y., WIPF D., YU Y., SNYDER J., TONG X.: Sparse-as-possible SVBRDF Acquisition. *ACM Trans. Graph. 35*, 6 (Nov. 2016), 189:1–189:12. 2

[Zha00] ZHANG Z.: A flexible new technique for camera calibration. *IEEE Transactions on pattern analysis and machine intelligence 22*, 11 (2000), 1330–1334. 3

[ZREB06] ZICKLER T., RAMAMOORTHI R., ENRIQUE S., BELHUMEUR P.: Reflectance sharing: predicting appearance from a sparse set of images of a known shape. *IEEE Transactions on Pattern Analysis and Machine Intelligence 28*, 8 (Aug. 2006), 1287–1302. 3

[ZWT13] ZHOU Z., WU Z., TAN P.: Multi-view photometric stereo with spatially varying isotropic materials. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2013), pp. 1482–1489. 3



blue damask fabric

green faux leather (one camera)

green faux leather (two cameras)

red velvet lamé (one camera)

red velvet lamé (two cameras)

wrapping paper (one camera)

wrapping paper (two cameras)

Figure 7: Example results showing the fitted svBRDF output. The top row of each material shows the results for one camera, while the bottom row (if available) shows the results for two cameras. Each row, from left to right: average color, ρ_d, ρ_s, α_x, α_y, and the normal offset map. Images have been cropped square and resized to fit.