On Joint Estimation of Pose, Geometry and svBRDF from a Handheld Scanner

Carolin Schmitt[1,2,*] Simon Donné[1,2,*] Gernot Riegler[3] Vladlen Koltun[3] Andreas Geiger[1,2]
[1]Max Planck Institute for Intelligent Systems, Tübingen
[2]University of Tübingen [3]Intel Intelligent Systems Lab
{firstname.lastname}@tue.mpg.de {firstname.lastname}@intel.com

Abstract

We propose a novel formulation for joint recovery of camera pose, object geometry and spatially-varying BRDF. The input to our approach is a sequence of RGB-D images captured by a mobile, hand-held scanner that actively illuminates the scene with point light sources. Compared to previous works that jointly estimate geometry and materials from a hand-held scanner, we formulate this problem using a single objective function that can be minimized using off-the-shelf gradient-based solvers. By integrating material clustering as a differentiable operation into the optimization process, we avoid pre-processing heuristics and demonstrate that our model is able to determine the correct number of specular materials independently. We provide a study on the importance of each component in our formulation and on the requirements of the initial geometry. We show that optimizing over the poses is crucial for accurately recovering fine details and that our approach naturally results in a semantically meaningful material segmentation.

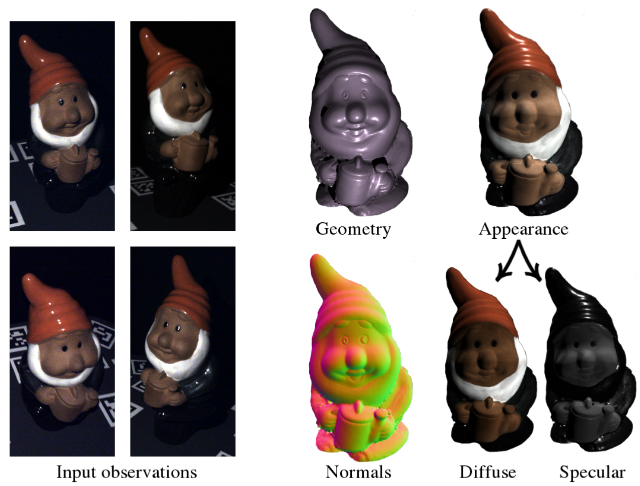


Figure 1: **Illustration.** Based on images captured from a handheld scanner with point light illumination, we jointly optimize for the camera poses, the surface geometry and spatially varying materials using a single objective function.

1. Introduction

Reconstructing the shape and appearance of objects is a long standing goal in computer vision and graphics with numerous applications ranging from telepresence to training embodied agents in photo-realistic environments. While novel depth sensing technology (e.g., Kinect) enabled large-scale 3D reconstructions [12, 61, 87], the level of realism provided is limited since physical light transport is not taken into account. As a consequence, material properties are not recovered and illumination effects such as specular reflections or shadows are merged into the texture component.

Material properties can be directly measured using dedicated light stages [26, 40, 49] or inferred from images by assuming known [15, 36, 65] or flat [3, 4, 29, 76] object geometry. However, most setups are either restricted to lab environments, planar geometries, or difficult to employ "in the wild" as they assume aligned 3D models or scans.

Ideally, object geometry and material properties are inferred jointly: a good model of light transport allows for recovering geometric detail using shading cues. An accurate shape model, in turn, facilitates the estimation of material properties. This is particularly relevant for shiny surfaces where small changes in the geometry greatly impact the appearance and location of specular reflections. Yet joint optimization of these quantities (shown in Fig. 1) is challenging.

Several works have addressed this problem by assuming multiple images from a static camera [16, 21, 23, 28, 105] which is impractical for mobile scanning applications. Only a few works consider the challenging problem of joint geometry and material estimation from a handheld device [20, 24, 57]. However, existing approaches assume known camera poses and leverage sophisticated pipelines, decomposing the problem into smaller problems using multiple decoupled objectives and optimization algorithms that treat geometry and materials separately. Furthermore the number of base materials must be provided and/or pre-processing is required to cluster the object surface accordingly.

* Joint first author with equal contribution.

In this work, we provide a novel formulation for this problem which does not rely on sophisticated pipelines or decoupled objective functions. However, we assume that the data was captured under known, non-static illumination with negligible ambient light. We make the following contributions: (1) We demonstrate that joint optimization of camera pose, object geometry and materials is possible using a single objective function and off-the-shelf gradient-based solvers. (2) We integrate material clustering as a differentiable operation into the optimization process by formulating non-local smoothness constraints. (3) Our approach automatically determines the number of specular base materials during the optimization process, leading to parsimonious and semantically meaningful material assignments. (4) We provide a study on the importance of each component in our formulation and a comparison to various baselines. (5) We provide our source code, dataset and reconstructed models publicly available at https://github.com/autonomousvision/handheld_svbrdf_geometry.

2. Related work

We now discuss the most related work on geometry, material as well as joint geometry and material estimation.

2.1. Geometry Estimation

Multi-View Stereo (MVS) reconstruction techniques [18, 38, 39, 77, 80, 84, 85] recover the 3D geometry of an object from multiple input images by matching feature correspondences across views or optimizing photo-consistency. As they ignore physical light transport, they cannot recover material properties. Furthermore, they are only able to recover geometry for surfaces which are sufficiently textured.

Shape from Shading (SfS) techniques exploit shading cues for reconstructing [27, 30, 72, 73, 99] or for refining 3D geometry [22, 48, 89, 104] from one or multiple images by relating surface normals to image intensities through Lambert's law. While early SfS approaches were restricted to objects made of a single Lambertian material, modern reincarnations of these models [6, 45, 62] are also able to infer non-Lambertian materials and lighting. Unfortunately, reconstructing geometry from a single image is a highly ill-posed problem, requiring strong assumptions about the surface geometry. Moreover, textured objects often cause ambiguities as intensity changes can be caused by changes in either surface orientation or surface albedo.

Photometric Stereo (PS) approaches [25, 63, 70, 71, 83, 88, 102] assume three or more images captured with a static camera while varying illumination or object pose [41, 82] to resolve the aforementioned ambiguities. In contrast to early PS approaches which often assumed orthographic cameras and distant light sources, newer works have considered the more practical setup of near light sources [42, 43, 74, 94] and perspective projection [53, 54, 68]. To handle non-Lambertian surfaces, robust error functions have been suggested [69, 75] and the problem has been formulated using specularity-invariant image ratios [10, 50–52]. The advantages of PS (accurate normals) and MVS (global geometry) have also been combined by integrating normals from PS and geometry from MVS [17, 33, 44, 47, 58, 64, 81, 96] into a single consistent reconstruction. However, many classical PS approaches are not capable of estimating material properties other than albedo and most PS approaches require a fixed camera which restricts their applicability to lab environments. In contrast, here we are interested in recovering shape and surface materials using a *handheld device*.

2.2. Material Estimation

Intrinsic Image Decomposition [6, 7, 11, 19] is the problem of decomposing an image into its material-dependent and light-dependent properties. However, only a small portion of the 3D physical process is captured by these models and strong regularizers must be exploited to solve the task. A more accurate description of the reflective properties of materials is provided by the Bidirectional Reflectance Distribution Function (BRDF) [59].

For **known 3D geometry**, the BRDF can be measured using specialized light stages or gantries [26, 40, 49, 60, 78]. While this setup leads to accurate reflectance estimates, it is typically expensive, stationary and only works for objects of limited size. In contrast, recent works have demonstrated that reflectance properties of flat surfaces can be acquired using an ordinary mobile phone [3, 4, 29, 76, 95]. While data collection is easy and practical, these techniques are designed for capturing flat texture surfaces and do not generalize to objects with more complex geometries.

More closely aligned with our goals are approaches that estimate parametric BRDF models for objects with known geometry based on sparse measurements of the BRDF space [15, 36, 55, 56, 65, 90–92, 97, 101]. While we also estimate a parametric BRDF model and assume only sparse measurements of the BRDF domain, we *jointly* optimize for camera pose, object geometry and material parameters. As our experiments show, joint optimization allows us to recover fine geometric structures (not present in the initial reconstruction) while at the same time improving material estimates compared to a sequential treatment of both tasks.

2.3. Joint Geometry and Material Estimation

Several works have addressed the problem of jointly inferring geometry and material. By integrating shading cues with multi-view constraints and an accurate model of materials and light transport, this approach has the potential to deliver the most accurate results. However, joint optimization of all relevant quantities is a challenging task.

Several works have considered extensions of the classic PS setting [1, 5, 8, 16, 21, 23, 28, 67, 93, 103, 105]. While some of these approaches consider multiple viewpoints and/or estimate spatially varying BRDFs, all of them require multiple images from the **same viewpoint** as input, i.e., they assume that the camera is on a tripod. While this would simplify matters, here we are interested in jointly estimating geometry and materials from a *handheld device* which can be used for scanning a wide range of objects outside the laboratory and which allows for obtaining more complete reconstructions by scanning objects from multiple viewpoints.

There exist only few works that consider the problem of joint geometry and material estimation from an **active handheld device**. Higo et al. [24] present a plane-sweeping approach combined with graph cuts for estimating albedo, normals and depth, followed by a post-processing step to integrate normal information into the depth and to remove outliers [58]. Georgoulis et al. [20] optimize an initial mesh computed via structure-from-motion and a data-driven BRDF model in an alternating fashion. They use k-means for clustering initial BRDF estimates into base materials and iteratively recompute the 3D geometry using the method of [58]. In similar spirit, Nam et al. [57] split the optimization problem into separate parts. They first cluster the material estimates using k-means, followed by an alternating optimization procedure which interleaves material, normal and geometry updates. The latter is updated using screened Poisson surface reconstruction [35] while materials are recovered using a separate objective.

The main contribution of our work is a simple and clean formulation of this problem: we demonstrate that geometry and materials can be inferred jointly using a *single objective function* optimized using standard gradient-based techniques. Our approach naturally allows for optimizing additional relevant quantities such as camera poses and integrates material clustering as a differentiable operation into the optimization process. Moreover, we demonstrate automatic model selection by determining the number of distinct material components as illustrated in Fig. 2.

3. Method

Let us assume a set of color images $\mathcal{I}_i : \mathbb{R}^2 \to \mathbb{R}^3$ captured from N different views $i \in \{1, \dots, N\}$. Without loss of generality, let us select $i = 1$ as the *reference view* based on which we will parameterize the surface geometry and materials as detailed below. Note that in our visualizations all observations are represented in this reference view.

Our goal is to jointly recover the camera poses, the geometry of the scene as well as the material properties in terms of a spatially varying Bidirectional Reflectance Distribution Function (svBRDF).

More formally, we wish to estimate the locations $\mathbf{x}_p = (x_p, y_p, z_p)^T$, surface normals $\mathbf{n}_p = (n_p^x, n_p^y, n_p^z)^T$, and svBRDF $f_p(\cdot)$ of a set of P surface points p, as well as the projective mappings $\pi_i : \mathbb{R}^3 \to \mathbb{R}^2$ that map a 3D point $\mathbf{x}_p \in \mathbb{R}^3$ into camera image i. We assume that each image is illuminated by exactly one point light source. Similar to prior works, we assume that global and ambient illumination effects are negligible.

3.1. Preliminaries

This section describes the parameterizations of our model.

Camera Representation: We use a perspective pinhole camera model and assume constant intrinsic camera parameters that can be estimated using established calibration procedures [100]. We also assume that all images have been undistorted and the vignetting has been removed. We therefore only optimize for the extrinsic parameters (i.e., rotation and translation) of each projective mapping $\pi_i : \mathbb{R}^3 \to \mathbb{R}^2$.

Geometry Representation: We define the surface points in terms of the depth map in the reference view $\mathcal{Z}_1 = \{z_p\}$, using p as the pixel/point index. Assuming a pinhole projection, the 3D location of surface point p is given by

$$\begin{aligned} \mathbf{x}_p &= \pi_1^{-1}(u_p, v_p, z_p) \\ &= \left(\frac{u_p - c_x}{f}, \frac{v_p - c_y}{f}, 1 \right) z_p \end{aligned} \tag{1}$$

where $[u_p, v_p]^T$ denotes the location of pixel p in the reference image $\mathcal{I}_c$, z_p is the depth at pixel p, π_1^{-1} is the inverse projection function and f, c_x, c_y denote its parameters.

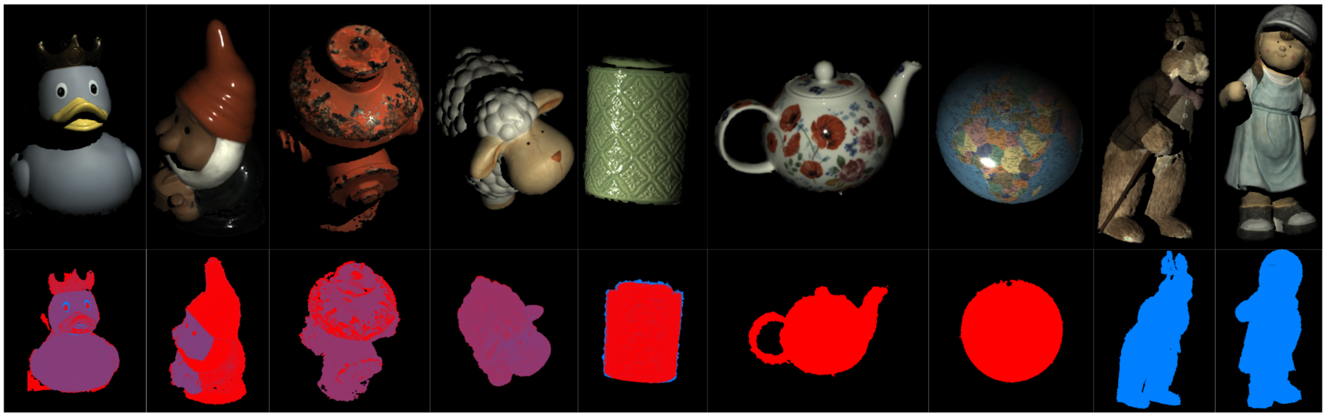
Normal Representation: We represent normals $\mathbf{n}_p$ as unit vectors. In every iteration of the gradient-based optimization, we estimate an angular change for this vector so that we avoid both the unit normal constraint and the gimbal lock problem.

svBRDF Representation: The svBRDF $f_p(\mathbf{n}_p, \boldsymbol{\omega}^{\text{in}}, \boldsymbol{\omega}^{\text{out}})$ models the fraction of light that is reflected from incoming light direction $\boldsymbol{\omega}^{\text{in}}$ to outgoing light direction $\boldsymbol{\omega}^{\text{out}}$ given the surface normal $\mathbf{n}_p$ at point p. We use a modified version of the Cook-Torrance microfacet BRDF model [13]

$$f_p(\mathbf{n}_p, \boldsymbol{\omega}^{\text{in}}, \boldsymbol{\omega}^{\text{out}}) = \mathbf{d}_p + \mathbf{s}_p \frac{D(r_p)\ G(\mathbf{n}_p, \boldsymbol{\omega}^{\text{in}}, \boldsymbol{\omega}^{\text{out}}, r_p)}{\pi(\mathbf{n}_p \cdot \boldsymbol{\omega}^{\text{in}})(\mathbf{n}_p \cdot \boldsymbol{\omega}^{\text{out}})} \tag{2}$$

where $D(\cdot)$ describes the microfacet slope distribution, $G(\cdot)$ is the geometric attenuation factor, and $\mathbf{d}_p \in \mathbb{R}^3$, $\mathbf{s}_p \in \mathbb{R}^3$ and $r_p \in \mathbb{R}$ denote diffuse albedo, specular albedo and surface roughness, respectively. We use Smith's function as implemented in Mitsuba [31] for $G(\cdot)$ and the GTR model of the Disney BRDF [9] for $D(\cdot)$. Following [57], we ignore the Fresnel effect which cannot be observed using a handheld setup.

As illustrated in Fig. 2, many objects can be modeled well with few specular material components [46] while the object texture is more complex. We thus allow the diffuse



Duck	Gnome	Hydrant	Sheep	Vase	Teapot	Globe	Rabbit	Girl
$(T=2)$	$(T=2)$	$(T=2)$	$(T=2)$	$(T=2)$	$(T=1)$	$(T=1)$	$(T=1)$	$(T=1)$

Figure 2: **Reconstructions and Estimated Material Assignments for Real Objects.** The teapot and the globe are very smooth and reflective objects whereas the girl and rabbit are near-Lambertian objects with rich geometry and a rough surface. The duck, the gnome and the hydrant are composed of both shiny and rough parts. While the duck, the gnome, the green vase and the sheep comprise mostly homogeneous colours, both the globe and the teapot have very detailed textures. The teapot is painted, which creates surface irregularities, while the globe is printed and therefore smooth. We show renderings (top) and specular base material segmentations (bottom) from our method; the number of bases T is automatically determined by model selection within our framework. We visualize the more specular materials in red.

albedo $\mathbf{d}_p$ to vary freely per pixel p, and model specular reflectance as a combination of T specular base materials

$$\binom{\mathbf{s}_p}{r_p} = \sum_{t=1}^{T} \alpha_p^t \binom{\mathbf{s}_t}{r_t} \tag{3}$$

with per-pixel BRDF weights $\alpha_p^t \in [0,1]$ and specular base materials $\{(\mathbf{s}_t, r_t)\}_{t=1}^T$. Note that this is in contrast to other representations [21,40] which linearly combine also the diffuse part, hence requiring more base materials to reach the same fidelity. We found that $T \leq 3$ specular bases are sufficient for almost all objects. In summary, our svBRDF is fully determined by $\{(\mathbf{s}_t, r_t)\}_{t=1}^T$ and $\{\mathbf{d}_p, \boldsymbol{\alpha}_p\}_{p=1}^P$.

3.2. Model

This section describes our objective function. Let $\mathcal{X} = \{\{(z_p, \mathbf{n}_p, f_p)\}_{p=1}^P, \{\pi_i\}_{i=2}^N\}$ denote the depth, normal and material for every pixel p in the reference view, as well as the projective mapping for each adjacent view. We formulate the following objective function

$$\mathcal{X}^* = \underset{\mathcal{X}}{\text{argmin}}\ \psi_\mathcal{P} + \psi_\mathcal{G} + \psi_\mathcal{D} + \psi_\mathcal{N} + \psi_\mathcal{M} \tag{4}$$

omitting the dependency on $\mathcal{X}$ and the relative weights between the individual terms. Our objective function is composed of five terms which encourage photoconsistency $\psi_\mathcal{P}$, geometric consistency $\psi_\mathcal{G}$, depth compatibility $\psi_\mathcal{D}$, normal smoothness $\psi_\mathcal{N}$ and material smoothness $\psi_\mathcal{M}$.

Photoconsistency: The photoconsistency term ensures that the prediction of our model matches the observation $\mathcal{I}_i$ for every image i and pixel p:

$$\psi_\mathcal{P}(\mathcal{X}) = \frac{1}{N}\sum_i \sum_p \| \varphi_p^i \left[\mathcal{I}_i(\pi_i(\mathbf{x}_p)) - \mathcal{R}_i(\mathbf{x}_p, \mathbf{n}_p, f_p)\right] \|_1 \tag{5}$$

Here, $\mathcal{R}_i$ denotes the *rendering operator* for image i which applies the rendering equation [34] to every pixel p. Assuming a single point light source, we obtain

$$\mathcal{R}_i(\mathbf{x}_p, \mathbf{n}_p, f_p) = f_p\left(\mathbf{n}_p, \boldsymbol{\omega}_i^{\text{in}}(\mathbf{x}_p), \boldsymbol{\omega}_i^{\text{out}}(\mathbf{x}_p)\right) \frac{a_i(\mathbf{x}_p)\, \mathbf{n}_p^T \boldsymbol{\omega}_i^{\text{in}}(\mathbf{x}_p)}{d_i(\mathbf{x}_p)^2} L \tag{6}$$

where $\boldsymbol{\omega}_i^{\text{in}}(\mathbf{x}_p)$ denotes the direction of the ray from the surface point $\mathbf{x}_p$ to the light source and $\boldsymbol{\omega}_i^{\text{out}}(\mathbf{x}_p)$ denotes the direction from $\mathbf{x}_p$ to the camera center. $a_i(\mathbf{x}_p)$ is the angle-dependent light attenuation which is determined through photometric calibration, $d_i(\mathbf{x}_p)$ is the distance between $\mathbf{x}_p$ and the light source and L denotes the radiant intensity of the light. Note that all terms depend on the image index i, as the location of the camera and the light source vary from frame to frame when recording with a handheld lightstage.

The visibility term φ_p^i in (5) disables occluded or shadowed observations i.e., we do not optimize for these regions. We set $\varphi_p^i = 1$ if surface point $\mathbf{x}_p$ is both visible in view i (i.e., no occluder between $\mathbf{x}_p$ and the i'th camera) *and* illuminated (e.g., no occluder between $\mathbf{x}_p$ and the point light), and $\varphi_p^i = 0$ otherwise. Note that for the reference view every pixel is visible, but not necessarily illuminated.

Geometric Consistency: We enforce consistency between depth $\{z_p\}$ and normals $\{\mathbf{n}_p\}$ by ensuring that the normal field integrates to the estimated depth map. We formulate this constraint by maximizing the inner product between the estimated normals $\{\mathbf{n}_p\}$ and the cross product of the surface tangents at $\{\mathbf{x}_p\}$:

$$\psi_{\mathcal{G}}(\mathcal{X}) = -\sum_p \vec{n}_p^T \left(\frac{\frac{\partial z_p}{\partial x} \times \frac{\partial z_p}{\partial y}}{\| \frac{\partial z_p}{\partial x} \times \frac{\partial z_p}{\partial y} \|_2} \right) \tag{7}$$

The surface tangent $\frac{\partial z_p}{\partial x}$ is given by

$$\frac{\partial z_p}{\partial x} \propto \left[1, 0, \vec{\nabla} \mathcal{Z}_1(\pi_1(\vec{x}_p))^T \left[f/z_p, 0 \right]^T \right]^T \tag{8}$$

where $\vec{\nabla} \mathcal{Z}_1(\pi_1(\vec{x}_p))$ denotes the gradient of the depth map, which we estimate using finite differences. We obtain a similar equation for $\frac{\partial z_p}{\partial y}$. See the supplement for details.

A valid question to raise is whether a separate treatment of depth and normals is necessary. An alternative formulation would consider consistency between depth and normals as a hard constraint, i.e., enforcing Equation (7) strictly, and optimizing only for depth. While reducing the number of parameters to be estimated, we found that such a representation is prone to local minima during optimization due to the complementary nature of the constraints (depth vs. normals/shading). Instead, using auxiliary normal variables and optimizing for both depth and normals using a soft coupling between them allows us to overcome these problems.

Depth Compatibility: The optional depth term allows for incorporating depth measurements $\mathcal{Z}_1$ in the reference view $i = 1$ by regularizing our estimates z_p against it:

$$\psi_{\mathcal{D}}(\mathcal{X}) = \sum_p \| z_p - \mathcal{Z}_1(u_p, v_p) \|_2^2 \tag{9}$$

Note that our model is able to significantly improve upon the initial coarse geometry provided by the structured light sensor by exploiting shading cues. However, as these cues are related to depth variations (i.e., normals) rather than absolute depth, they do not fully constrain the 3D shape of the object. Our experiments demonstrate that combining complementary depth and shading cues yields reconstructions which are both locally detailed and globally consistent.

Normal Smoothness: We apply a standard smoothness regularizer to the normals of adjacent pixels $p \sim q$

$$\psi_{\mathcal{G}}(\mathcal{X}) = \sum_{p \sim q} \| \mathbf{n}_p - \mathbf{n}_q \|_2^2 \tag{10}$$

in order to encourage smooth surfaces.

Material Smoothness: We only observe specular BRDF components for a minority of pixels that actually observe a specular highlight in at least one of their measurements. We therefore introduce a non-local material regularizer which propagates specular behavior across image regions of similar appearance. Assuming that nearby pixels with similar diffuse behavior also exhibit similar specular behavior, we formulate this term by penalizing deviation of the material weights wrt. a bilaterally smoothed version of themselves

$$\begin{aligned} \psi_{\mathcal{M}}(\mathcal{X}) = \sum_p & \left\| \boldsymbol{\alpha}_p - \frac{\sum_q \boldsymbol{\alpha}_q \, w_q k_{p,q}}{\sum_q w_q k_{p,q}} \right\|_1 \\ & - \sum_p \left\| \boldsymbol{\alpha}_p - \frac{1}{P} \sum_q \boldsymbol{\alpha}_q \right\|_1 \end{aligned} \tag{11}$$

using a Gaussian kernel $k_{p,q}(\mathbf{d}_p, \mathbf{d}_q)$ with 3D location $\mathbf{x}$ and diffuse albedo $\mathbf{d}$ at pixels p and q as features:

$$k_{p,q} = \exp\left(-\frac{(\mathbf{x}_p - \mathbf{x}_q)_2^2}{2\sigma_1^2} - \frac{(\mathbf{d}_p - \mathbf{d}_q)_2^2}{2\sigma_2^2} \right) \tag{12}$$

As the only informative regions are those that potentially observe a highlight, the weights $w_q = \max_i \operatorname{acos}^{-1}(\mathbf{n}_q \cdot \mathbf{h}_q^i)$ indicate whether pixel q was ever observed close to perfect mirror reflection. This is determined by the normal $\mathbf{n}_q$ and the half-vector $\mathbf{h}_p^i$ (i.e., the bisector between $\boldsymbol{\omega}^{in}$ and $\boldsymbol{\omega}^{out}$) for each view i. We use the permutohedral lattice [2] to efficiently evaluate the bilateral filter.

The second term in (11) encourages material sparsity by maximizing the distance to the average BRDF weights where P denotes the total number of surface points/pixels.

3.3. Optimization

We now discuss the parameter initialization and how we minimize our objective function (4) with respect to $\mathcal{X}$.

Initial Poses: The camera poses can be either initialized using classical SfM pipelines such as COLMAP [77, 79] or using a set of fiducial markers. As SfM approaches fail in the presence of textureless surfaces, we use a small set of AprilTags [86] attached to the table supporting the object of interest. As evidenced by our experiments, the poses estimated using fiducial markers are not accurate enough to model pixel-accurate light transport. We demonstrate that geometry and materials can be significantly improved by jointly refining the initial camera poses.

Initial Depth: The initial depth map $\mathcal{Z} = \{z_p\}$ can be obtained using active or passive stereo, or the visual hull of the object. As we do not assume textured objects and silhouettes can be difficult to extract in the presence of dark materials, we use active stereo with a Kinect-like dot pattern projector for estimating $\mathcal{Z}$. More specifically, we estimate a depth map for each of the N views, integrate them using volumetric fusion [14] and project the resulting mesh back to the reference view.

Observation Crop Reconstruction

Figure 3: **Super-Resolution and Denoising.** Blurred and noisy input images (Observation and Crop) get denoised and sharpened by our method (Reconstruction).

Initial Normals and Albedo: Assuming a Lambertian scene, normals and albedo can be recovered in closed form. We follow the approach of Higo et al. [24] and use RANSAC to reject outliers due to specularities.

Initial Specular BRDF Parameters and Weights: We initialize each pixel in the scene as a uniform mix of all base materials. To diversify the initial base materials, we initialize the specular base components $\mathbf{s}_t$ differently, and set each base roughness r_t to 0.1.

Model Selection: We perform model selection by optimizing for multiple numbers of specular base materials $T \in \{1, 2, 3\}$, choosing that with the smallest photometric error while adding a small MDL penalty (linear in T).

Implementation: We jointly optimize over $\mathcal{X}$, using ADAM [37] and PyTorch [66], see supplement for details.

4. Experimental Evaluation

In order to evaluate our method quantitatively and qualitatively, we capture several real objects using a custom-built rig with active illumination. Reconstructions of these objects are shown in Fig. 2. We scanned the objects with an Artec Spider[1] to obtain ground truth geometry.

We first briefly describe our hardware and data capture procedure. After introducing the metrics, both geometric and photometric, we provide an ablation study in terms of the individual components of our model. Finally, we compare our approach with several competitive baselines.

4.1. Evaluation Protocol

Hardware: Our custom-built handheld sensor rig is shown in Fig. 4. While we use multiple light sources for a dense sampling of the BRDF, our framework and code is directly applicable to any number of lights. We calibrate the camera and depth sensor regarding their intrinsics and extrinsics as well as vignetting effects. We also calibrate the location of the light sources relative to the camera as well as their angular attenuation behavior and radiant intensities. Due to space limitations, we provide more details about our system in the supplement.

[1] https://www.artec3d.com/portable-3d-scanners/artec-spider

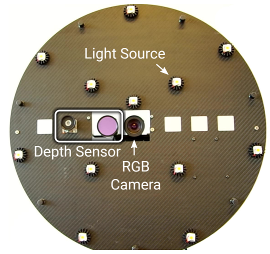


Figure 4: **Sensor Rig.** Our setup comprises a Kinect-like active depth sensor, a global shutter RGB camera and 12 point light sources (high-power LEDs) surrounding the camera in two circles (with radii 10 cm and 25 cm).

Data Capture: The objects are placed on a table with AprilTags [86] for tracking the sensor position. We assume that the ambient light is negligible and capture videos of each object by moving the handheld sensor around the object. Given each reference view, we select 45 views within a viewing cone of 30 degrees by maximizing the minimum pairwise distance; no two views are ever close together. These views are then split into 40 training and 5 held-out test views. While a handheld setup is challenging due to the trade-off between motion blur and image noise, our experiments demonstrate that our method is capable to super-resolve and denoise fine textures while simultaneously rejecting blurry observations, see Fig. 3.

Evaluation Metrics: We evaluate the estimated structure $\{\mathbf{x}_p\}$ wrt. the Artec Spider scan. The ground truth scan is first roughly aligned by hand and subsequently finetuned using dense image-based alignment wrt. depth and normal errors. We evaluate geometric accuracy by using the average point-to-mesh distance for all reconstructed points as in [32]. To evaluate surface normals, we calculate the average angular error (AAE) between the predicted normal $\mathbf{n}_p$ and the normal of the closest point in the ground truth scan. To quantify photometric reconstruction quality, we calculate the photoconsistency term in Eq. (5) for the test views.

4.2. Ablation Study

In this section we demonstrate the need to optimize over the camera poses and discuss the effect of specifically the geometric consistency and the material smoothness terms. Finally, we investigate the impact of the number of views on the photometric and geometric error. Additional results are provided in the supplement.

Pose Optimization: Disambiguating geometric properties from material is a major challenge. We found that optimizing the poses jointly with the other parameters is crucial for this, in particular when working with a handheld scanner. Fig. 5 shows that inaccurate poses cause a significant contamination of the geometry with texture information. This is even more crucial when estimating specularities: misalignment causes highlights to be inconsistent with the geometry and therefore difficult to recover.

Photometric Test Error	Overall	Specular	Non-Specular
Fixed Poses	1.210	3.349	1.151
Full Model	**1.138**	**3.243**	**1.081**

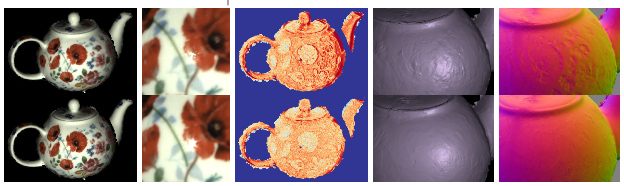


Figure 5: **Pose Optimization.** Compared to using the input poses (top), optimizing the poses (bottom) improves reconstructions, both quantitatively and qualitatively. The photometric error is reported for regions with and without specular highlights. See the supplement for more details.

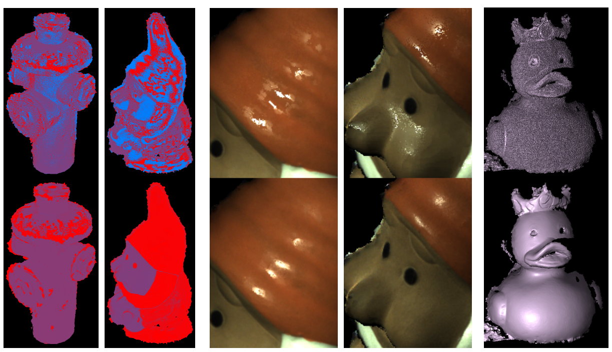


Figure 6: **Loss Regularizers.** Without the regularization (top), the appearance is inconsistent within homogeneous areas of the object. Using the regularization losses (bottom), we are able to propagate the information and successfully generalize to new illumination conditions on the test set.

Material Segmentation: Decomposing the appearance of the object into its individual materials is an integral element of our approach. Our material smoothness term Eq. (11) propagates material information over large areas of the image. This is essential as we otherwise only obtain sparse measurements of the BRDF at each pixel. It leads to semantically meaningfull segmentations, as illustrated in Fig. 2, as well as more successfull generalization, as shown in Fig. 6a.

Geometric Consistency: Splitting up the depth and normals into separate optimization variables yields a better behaved optimization problem, but coupling depth and normals proves crucial for consistent results. Even though the photometric term provides some constraints for the depth at each pixel, Fig. 6b shows that omitting the geometric consistency term results in high-frequency structure artifacts.

Number of Input Views: Our goal is to estimate the spatially varying BRDF but we only observe a very sparse set

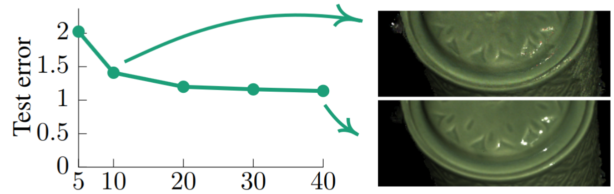


Figure 7: **Number of Input Views.** The photometric test error (blue) degrades gracefully with decreasing number of observations. As expected, the quality of the highlights is most affected by a small number of views.

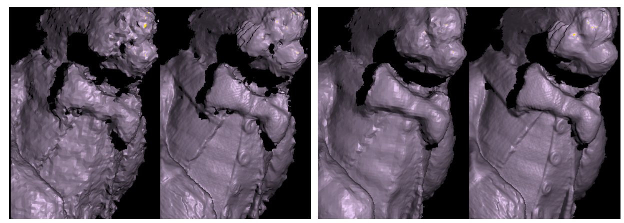


Figure 8: **Geometry Refinement.** The final geometry estimate (right) is largely unaffected by the initialization (left). Details are recovered, even from a very coarse initialization.

of samples for each surface point p. Reducing the number of images exacerbates this problem, as shown in Fig. 7. We see that our method degrades gracefully, with reasonable results even for using only 10 input images.

We also evaluate the robustness of our method wrt. the initial geometry by reducing the number of depth maps fused for initialization. As Fig. 8 shows, our method is able to recover from inaccurate depth initialization and achieves similar quality reconstructions even when initializing from only 5 depth maps. By construction, our model does not recover geometry that is absent in the initial estimate.

4.3. Comparison to Existing Approaches

Similar to us, Higo et al. [24] use a handheld scanner for estimating depth, normals and material using a 2.5D representation. Unlike us, they treat specular highlights, shadows and occlusions as outliers using RANSAC. Georgoulis et al. [20] and Nam et al. [57] also estimate structure and normals, explicitly modeling non-Lambertian materials. But due to the nature of their pipelines, they are restricted to a disjoint optimization procedure and update geometry and materials in alternation. It is important to note that the baselines expect the camera positions to be known accurately. So for the baselines we first refine the poses using SfM [77]. Unfortunately, none of the existing works provide code. We have re-implemented the approach of Higo et al. [24] as baseline. To investigate the benefits of joint optimization, we implemented a *disjoint* variant of our method that alternates between geometry and material

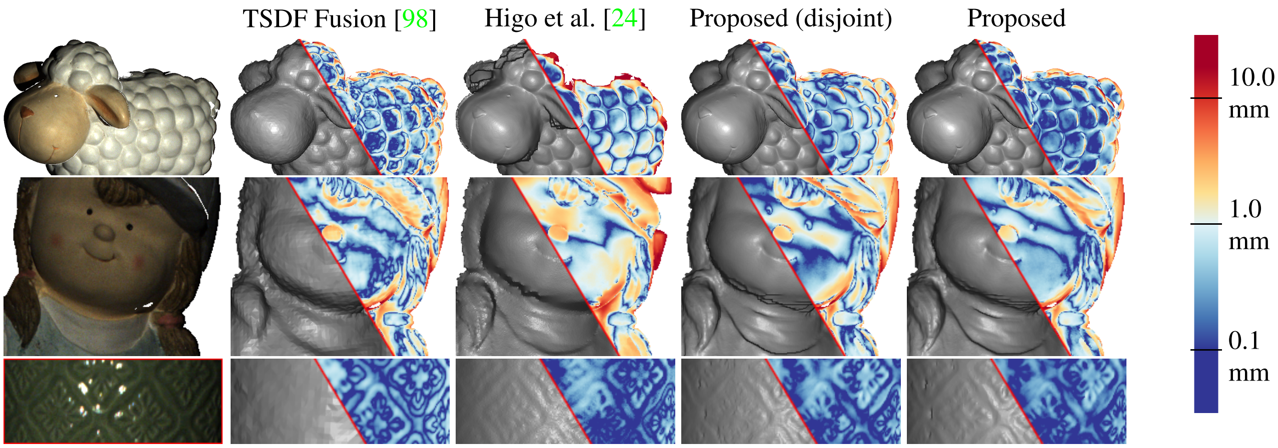


Figure 9: **Qualitative Geometry Comparison.** We show, for each object, the rendered depth map (shaded based on estimated surface normals rather than the estimated normals $\mathbf{n}_p$) and the color-coded depth error wrt. the Artec Spider ground truth. We observe that the photometric approaches recover far more details than the purely geometric TSDF fusion. The resulting structure for Higo et al. [24] is rather noisy, while the disjoint version of our proposed approach is not as successful at disambiguating texture and geometry for very fine details. We refer to the supplement for additional results.

		duck	pineapple	girl	gnome	sheep	hydrant	rabbit
AEA	TSDF Fusion [98]	0.81	1.24	1.11	0.73	0.79	1.35	**2.16**
	Higo et al. [24]	2.65	1.05	1.59	1.60	2.09	1.81	2.85
	Proposed (disjoint)	0.81	**1.00**	1.06	0.65	0.64	1.18	2.25
	Proposed	**0.80**	**1.00**	**1.00**	**0.64**	**0.57**	**1.16**	**2.16**
AAE	TSDF Fusion [98]	6.75	12.09	11.40	7.64	8.38	11.67	24.42
	Higo et al. [24]	7.77	10.62	11.30	9.24	8.65	15.27	27.67
	Proposed (disjoint)	6.01	9.13	9.81	6.41	6.66	9.59	**23.01**
	Proposed	**5.17**	**8.98**	**8.73**	**5.74**	**5.60**	**8.50**	23.50

Table 1: **Quantitative Geometry Comparison.** We report both the average euclidean accuracy and average angular error, as discussed in Section 4.1. Please refer to the supplement for a more detailed table.

updates. We also evaluate the improvement over the initial TSDF fusion using the implementation of Zeng et al. [98].

Our experimental evaluation[2] is shown in Fig. 9 and Table 1. We see that TSDF fusion, a purely geometric approach, reconstructs the general surface well but misses fine details. Their spatial regularizer helps Higo et al. [24] to achieve reasonable reconstructions, which are however strongly affected by the noisy, unregularized, normal estimates. Additionally, the RANSAC approach to shadow handling results in artifacts around depth discontinuities.

Both the *joint* and *disjoint* versions of our approach reconstruct the scene more accurately than the baselines, but the joint approach consistently obtains better reconstruction accuracy given a fixed computational budget.

Glossy black materials, such as the eyes of the duck (Fig. 6b) or the rabbit (Fig. 2), remain a significant challenge. For such materials, the signal-to-noise of the diffuse component is low and the signal from specular highlights is very sparse so that neither the photoconsistency nor the depth compatibility term constrain the solution correctly.

[2]We do not include results of traditional MVS techniques here, as they fail for textureless objects. We include them in the supplement.

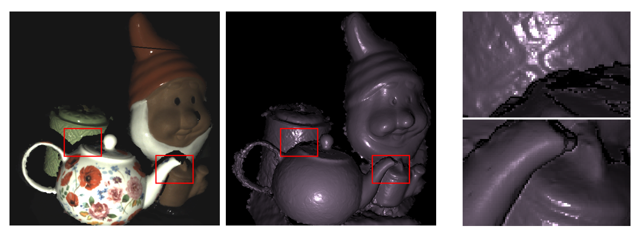


Figure 10: **A strongly Non-Convex Scene.** Due to our explicit shadow and occlusion model, our approach is also applicable to the reconstruction of more complex scenes.

Finally, Fig. 10 illustrates that our approach is able to also handle strongly non-convex scenes whose shadows and occlusions often cause issues for existing methods.

5. Conclusion

We have proposed a practical approach to estimating geometry and materials from a handheld sensor. Accurate camera poses are crucial to this task, but are not readily available. To tackle this problem, we propose a novel formulation which enables joint optimization of poses, geometry and materials using a single objective. Our approach recovers accurate geometry and material properties, and produces a semantically meaningful set of material weights.

Acknowledgements: This work was supported by the Intel Network on Intelligent Systems (NIS). We thank Lars Mescheder and Michal Rolínek for their feedback and the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Carolin Schmitt.

References

[1] Jens Ackermann, Martin Ritz, André Stork, and Michael Goesele. Removing the example from example-based photometric stereo. In *ECCV Workshops*, 2010. 2

[2] Andrew Adams, Jongmin Baek, and Myers Abraham Davis. Fast high-dimensional filtering using the permutohedral lattice. *Computer Graphics Forum*, 29(2):753–762, 2010. 5

[3] Miika Aittala, Tim Weyrich, and Jaakko Lehtinen. Two-shot SVBRDF capture for stationary materials. *ACM Trans. on Graphics*, 34(4):110:1–110:13, 2015. 1, 2

[4] Rachel A. Albert, Dorian Yao Chan, Dan B. Goldman, and James F. O'Brien. Approximate svbrdf estimation from mobile phone video. In *EUROGRAPHICS*, 2018. 1, 2

[5] Neil Gordon Alldrin, Todd E. Zickler, and David J. Kriegman. Photometric stereo with non-parametric and spatially-varying reflectance. In *CVPR*, 2008. 2

[6] Jonathan T. Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *PAMI*, 37(8):1670–1687, 2015. 2

[7] H.G. Barrow. Recovering intrinsic scene characteristics from images. *CVS*, pages 3–26, 1978. 2

[8] Neil Birkbeck, Dana Cobzas, Peter F. Sturm, and Martin Jägersand. Variational shape and reflectance estimation under changing light and viewpoints. In *ECCV*, 2006. 2

[9] Brent Burley. Physically-based shading at Disney. Technical report, Walt Disney Animation Studios, 2012. 3

[10] Manmohan Krishna Chandraker, Jiamin Bai, and Ravi Ramamoorthi. A theory of differential photometric stereo for unknown isotropic brdfs. In *CVPR*, 2011. 2

[11] Qifeng Chen and Vladlen Koltun. A simple model for intrinsic image decomposition with depth cues. In *ICCV*, 2013. 2

[12] Sungjoon Choi, Qian-Yi Zhou, and Vladlen Koltun. Robust reconstruction of indoor scenes. In *CVPR*, 2015. 1

[13] Robert L. Cook and Kenneth E. Torrance. A reflectance model for computer graphics. *ACM Trans. on Graphics*, 1(1):7–24, 1982. 3

[14] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *ACM Trans. on Graphics*, 1996. 5

[15] Yue Dong, Guojun Chen, Pieter Peers, Jiawan Zhang, and Xin Tong. Appearance-from-motion: recovering spatially varying surface reflectance under unknown lighting. *ACM Trans. on Graphics*, 33(6):193:1–193:12, 2014. 1, 2

[16] Carlos Hernández Esteban, George Vogiatzis, and Roberto Cipolla. Multiview photometric stereo. *PAMI*, 30(3):548–554, 2008. 1, 2

[17] Hao Fan, Lin Qi, Junyu Dong, Gongfa Li, and Hui Yu. Dynamic 3d surface reconstruction using a hand-held camera. In *IECON*, 2018. 2

[18] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multi-view stereopsis. *PAMI*, 32(8):1362–1376, 2010. 2

[19] Peter V. Gehler, Carsten Rother, Martin Kiefel, Lumin Zhang, and Bernhard Schölkopf. Recovering intrinsic images with a global sparsity prior on reflectance. In *NIPS*, pages 765–773, 2011. 2

[20] Stamatios Georgoulis, Marc Proesmans, and Luc J. Van Gool. Tackling shapes and brdfs head-on. In *3DV*, 2014. 1, 3, 7

[21] Dan B. Goldman, Brian Curless, Aaron Hertzmann, and Steven M. Seitz. Shape and spatially-varying brdfs from photometric stereo. *PAMI*, 32(6):1060–1071, 2010. 1, 2, 4

[22] Bjoern Haefner, Yvain Quéau, Thomas Möllenhoff, and Daniel Cremers. Fight ill-posedness with ill-posedness: Single-shot variational depth super-resolution from shading. In *CVPR*, 2018. 2

[23] Aaron Hertzmann and Steven M. Seitz. Example-based photometric stereo: Shape reconstruction with general, varying brdfs. *PAMI*, 27(8):1254–1264, 2005. 1, 2

[24] Tomoaki Higo, Yasuyuki Matsushita, Neel Joshi, and Katsushi Ikeuchi. A hand-held photometric stereo camera for 3-d modeling. In *ICCV*, 2009. 1, 3, 6, 7, 8

[25] Michael Holroyd, Jason Lawrence, Greg Humphreys, and Todd E. Zickler. A photometric approach for estimating normals and tangents. *ACM Trans. on Graphics*, 27(5):133:1–133:9, 2008. 2

[26] Michael Holroyd, Jason Lawrence, and Todd E. Zickler. A coaxial optical scanner for synchronous acquisition of 3d geometry and surface reflectance. *ACM Trans. on Graphics*, 29(4):99:1–99:12, 2010. 1, 2

[27] B. K.P. Horn. Shape from shading: A method for obtaining the shape of a smooth opaque object from one view bibtex. Technical report, 1970. 2

[28] Z. Hui and A. C. Sankaranarayanan. Shape and spatially-varying reflectance estimation from virtual exemplars. *PAMI*, 2017. 1, 2

[29] Z. Hui, K. Sunkavalli, J. Lee, S. Hadap, J. Wang, and A. C. Sankaranarayanan. Reflectance capture using univariate sampling of brdfs. In *ICCV*, 2017. 1, 2

[30] Katsushi Ikeuchi and Berthold K. P. Horn. Numerical shape from shading and occluding boundaries. *AI*, 17(1-3):141–184, 1981. 2

[31] Wenzel Jakob. Mitsuba renderer, 2010. http://www.mitsuba-renderer.org. 3

[32] Rasmus Ramsbøl Jensen, Anders Lindbjerg Dahl, George Vogiatzis, Engil Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *CVPR*, 2014. 6

[33] Neel Joshi and David J. Kriegman. Shape from varying illumination and viewpoint. In *ICCV*, 2007. 2

[34] James T. Kajiya. The rendering equation. In *ACM Trans. on Graphics*, 1986. 4

[35] Michael M. Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM Trans. on Graphics*, 32(3):29, 2013. 3

[36] Kihwan Kim, Jinwei Gu, Stephen Tyree, Pavlo Molchanov, Matthias Nießner, and Jan Kautz. A lightweight approach for on-the-fly reflectance estimation. In *ICCV*, 2017. 1, 2

[37] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[38] Vladimir Kolmogorov and Ramin Zabih. Multi-camera scene reconstruction via graph cuts. In *ECCV*, 2002. 2

[39] Florent Lafarge, Renaud Keriven, Mathieu Bredif, and Hoang-Hiep Vu. A hybrid multiview stereo algorithm for modeling urban scenes. *PAMI*, 35(1):5–17, 2013. 2

[40] Hendrik P. A. Lensch, Jan Kautz, Michael Goesele, Wolfgang Heidrich, and Hans-Peter Seidel. Image-based reconstruction of spatial appearance and geometric detail. *ACM Trans. on Graphics*, 22(2):234–257, 2003. 1, 2, 4

[41] Jongwoo Lim, Jeffrey Ho, Ming-Hsuan Yang, and David J. Kriegman. Passive photometric stereo from motion. In *ICCV*, 2005. 2

[42] Chao Liu, Srinivasa G. Narasimhan, and Artur W. Dubrawski. Near-light photometric stereo using circularly placed point light sources. 2018. 2

[43] Fotios Logothetis, Roberto Mecca, and Roberto Cipolla. Semi-calibrated near field photometric stereo. In *CVPR*, 2017. 2

[44] Fotios Logothetis, Roberto Mecca, and Roberto Cipolla. A differential volumetric approach to multi-view photometric stereo. In *ICCV*, 2019. 2

[45] Stephen Lombardi and Ko Nishino. Single image multimaterial estimation. In *CVPR*, 2012. 2

[46] Stephen Lombardi and Ko Nishino. Radiometric scene decomposition: Scene reflectance, illumination, and geometry from RGB-D images. In *3DV*, 2016. 3

[47] Zheng Lu, Yu-Wing Tai, Moshe Ben-Ezra, and Michael S. Brown. A framework for ultra high resolution 3d imaging. In *CVPR*, 2010. 2

[48] Robert Maier, Kihwan Kim, Daniel Cremers, Jan Kautz, and Matthias Nießner. Intrinsic3d: High-quality 3d reconstruction by joint appearance and geometry optimization with spatially-varying lighting. In *ICCV*, pages 3133–3141, 2017. 2

[49] Wojciech Matusik, Hanspeter Pfister, Matt Brand, and Leonard McMillan. A data-driven reflectance model. *ACM Trans. on Graphics*, 22(3):759–769, July 2003. 1, 2

[50] Roberto Mecca and Yvain Quéau. Unifying diffuse and specular reflections for the photometric stereo problem. In *WACV*, 2016. 2

[51] Roberto Mecca, Yvain Quéau, Fotios Logothetis, and Roberto Cipolla. A single-lobe photometric stereo approach for heterogeneous material. *SIAM*, 9(4):1858–1888, 2016. 2

[52] Roberto Mecca, Emanuele Rodolà, and Daniel Cremers. Realistic photometric stereo using partial differential irradiance equation ratios. *Computers & Graphics*, 51:8–16, 2015. 2

[53] Roberto Mecca, Ariel Tankus, Aaron Wetzler, and Alfred M. Bruckstein. A direct differential approach to photometric stereo with perspective viewing. *SIAM*, 7(2):579–612, 2014. 2

[54] Roberto Mecca, Aaron Wetzler, Alfred M. Bruckstein, and Ron Kimmel. Near field photometric stereo with point light sources. *SIAM*, 7(4):2732–2770, 2014. 2

[55] Jean Mélou, Yvain Quéau, Jean-Denis Durou, Fabien Castan, and Daniel Cremers. Beyond multi-view stereo: Shading-reflectance decomposition. In *SSVM*, pages 694–705, 2017. 2

[56] Jean Mélou, Yvain Quéau, Jean-Denis Durou, Fabien Castan, and Daniel Cremers. Variational reflectance estimation from multi-view images. *JMIV*, 60(9):1527–1546, 2018. 2

[57] Giljoo Nam, Joo Ho Lee, Diego Gutierrez, and Min H. Kim. Practical SVBRDF acquisition of 3d objects with unstructured flash photography. *ACM Trans. on Graphics*, 37(6):267:1–267:12, 2018. 1, 3, 7

[58] Diego Nehab, Szymon Rusinkiewicz, James Davis, and Ravi Ramamoorthi. Efficiently combining positions and normals for precise 3d geometry. *ACM Trans. on Graphics*, 24(3):536–543, 2005. 2, 3

[59] F. E. Nicodemus, J. C. Richmond, J. J. Hsia, I. W. Ginsberg, and T. Limperis. Geometrical considerations and nomenclature for reflectance. In *Radiometry*, chapter Geometrical Considerations and Nomenclature for Reflectance, pages 94–145. Jones and Bartlett Publishers, Inc., 1992. 2

[60] Jannik Boll Nielsen, Henrik Wann Jensen, and Ravi Ramamoorthi. On optimal, minimal BRDF sampling for reflectance acquisition. *ACM Trans. on Graphics*, 34(6):186:1–186:11, 2015. 2

[61] M. Nießner, M. Zollhöfer, S. Izadi, and M. Stamminger. Real-time 3d reconstruction at scale using voxel hashing. In *ACM Trans. on Graphics*, 2013. 1

[62] Geoffrey Oxholm and Ko Nishino. Multiview shape and reflectance from natural illumination. In *CVPR*, 2014. 2

[63] Thoma Papadhimitri and Paolo Favaro. Uncalibrated near-light photometric stereo. In *BMVC*, 2014. 2

[64] Jaesik Park, Sudipta N. Sinha, Yasuyuki Matsushita, Yu-Wing Tai, and In So Kweon. Robust multiview photometric stereo using planar mesh parameterization. *PAMI*, 39(8):1591–1604, 2017. 2

[65] Jeong Joon Park, Richard A. Newcombe, and Steven M. Seitz. Surface light field fusion. In *3DV*, 2018. 1, 2

[66] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *NIPS Workshops*, 2017. 6

[67] Songyou Peng, Bjoern Haefner, Yvain Quéau, and Daniel Cremers. Depth super-resolution meets uncalibrated photometric stereo. In *ICCV Workshops*, 2017. 2

[68] Yvain Quéau, Bastien Durix, Tao Wu, Daniel Cremers, François Lauze, and Jean-Denis Durou. Led-based photometric stereo: Modeling, calibration and numerical solution. *JMIV*, 60(3):313–340, 2018. 2

[69] Yvain Quéau, François Lauze, and Jean-Denis Durou. A l^1 -tv algorithm for robust perspective photometric stereo with spatially-varying lightings. In *SSVM*, 2015. 2

[70] Yvain Quéau, François Lauze, and Jean-Denis Durou. Solving uncalibrated photometric stereo using total variation. *JMIV*, 52(1):87–107, 2015. 2

[71] Yvain Quéau, Roberto Mecca, and Jean-Denis Durou. Unbiased photometric stereo for colored surfaces: A variational approach. In *CVPR*, 2016. 2

[72] Yvain Quéau, Jean Mélou, Fabien Castan, Daniel Cremers, and Jean-Denis Durou. A variational approach to shape-from-shading under natural illumination. In *EMMCVPR*, 2017. 2

[73] Yvain Quéau, Jean Mélou, Jean-Denis Durou, and Daniel Cremers. Dense multi-view 3d-reconstruction without dense correspondences. *arXiv.org*, 1704.00337, 2017. 2

[74] Yvain Quéau, Tao Wu, and Daniel Cremers. Semi-calibrated near-light photometric stereo. In *SSVM*, 2017. 2

[75] Yvain Quéau, Tao Wu, François Lauze, Jean-Denis Durou, and Daniel Cremers. A non-convex variational approach to photometric stereo under inaccurate lighting. In *CVPR*, 2017. 2

[76] Jérémy Rivière, Pieter Peers, and Abhijeet Ghosh. Mobile surface reflectometry. *Computer Graphics Forum*, 35(1):191–202, 2016. 1, 2

[77] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*, 2016. 2, 5, 7

[78] Christopher Schwartz, Ralf Sarlette, Michael Weinmann, and Reinhard Klein. DOME II: A parallelized BTF acquisition system. In *EUROGRAPHICS*, 2013. 2

[79] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 5

[80] S.M. Seitz and C.R. Dyer. Photorealistic scene reconstruction by voxel coloring. In *CVPR*, 1997. 2

[81] Boxin Shi, Kenji Inose, Yasuyuki Matsushita, Ping Tan, Sai-Kit Yeung, and Katsushi Ikeuchi. Photometric stereo using internet images. In *3DV*, 2014. 2

[82] Denis Simakov, Darya Frolova, and Ronen Basri. Dense shape reconstruction of a moving object under arbitrary, unknown lighting. In *ICCV*, 2003. 2

[83] Borom Tunwattanapong, Graham Fyffe, Paul Graham, Jay Busch, Xueming Yu, Abhijeet Ghosh, and Paul E. Debevec. Acquiring reflectance and shape from continuous spherical harmonic illumination. *ACM Trans. on Graphics*, 32(4):109:1–109:12, 2013. 2

[84] Ali Osman Ulusoy, Andreas Geiger, and Michael J. Black. Towards probabilistic volumetric reconstruction using ray potentials. In *3DV*, 2015. 2

[85] George Vogiatzis, Philip H. S. Torr, and Roberto Cipolla. Multi-view stereo via volumetric graph-cuts. In *CVPR*, 2005. 2

[86] John Wang and Edwin Olson. AprilTag 2: Efficient and robust fiducial detection. In *IROS*, October 2016. 5, 6

[87] Thomas Whelan, Stefan Leutenegger, Renato F. Salas-Moreno, Ben Glocker, and Andrew J. Davison. Elasticfusion: Dense SLAM without A pose graph. In *RSS*, 2015. 1

[88] Robert J Woodham. Photometric method for determining surface orientation from multiple images. *OE*, 19(1):191139, 1980. 2

[89] Chenglei Wu, Michael Zollhöfer, Matthias Nießner, Marc Stamminger, Shahram Izadi, and Christian Theobalt. Realtime shading-based refinement for consumer depth cameras. In *ACM Trans. on Graphics*, 2014. 2

[90] Hongzhi Wu, Zhaotian Wang, and Kun Zhou. Simultaneous localization and appearance estimation with a consumer RGB-D camera. *VCG*, 22(8):2012–2023, 2016. 2

[91] Hongzhi Wu and Kun Zhou. Appfusion: Interactive appearance acquisition using a kinect sensor. *Computer Graphics Forum*, 34(6):289–298, 2015. 2

[92] Zhe Wu, Sai-Kit Yeung, and Ping Tan. Towards building an RGBD-M scanner. *arXiv.org*, 1603.03875, 2016. 2

[93] Rui Xia, Yue Dong, Pieter Peers, and Xin Tong. Recovering shape and spatially-varying surface reflectance under unknown illumination. *ACM Trans. on Graphics*, 35(6):187:1–187:12, 2016. 2

[94] Wuyuan Xie, Chengkai Dai, and Charlie C. L. Wang. Photometric stereo with near point lighting: A solution by mesh deformation. In *CVPR*, 2015. 2

[95] Zexiang Xu, Jannik Boll Nielsen, Jiyang Yu, Henrik Wann Jensen, and Ravi Ramamoorthi. Minimal BRDF sampling for two-shot near-field reflectance acquisition. *ACM Trans. on Graphics*, 35(6):188:1–188:12, 2016. 2

[96] Yusuke Yoshiyasu and Nobutoshi Yamazaki. Topology-adaptive multi-view photometric stereo. In *CVPR*, 2011. 2

[97] Yizhou Yu, Paul E. Debevec, Jitendra Malik, and Tim Hawkins. Inverse global illumination: Recovering reflectance models of real scenes from photographs. In *ACM Trans. on Graphics*, pages 215–224, 1999. 2

[98] Andy Zeng, Shuran Song, Matthias Nießner, Matthew Fisher, Jianxiong Xiao, and Thomas Funkhouser. 3dmatch: Learning local geometric descriptors from rgb-d reconstructions. In *CVPR*, 2017. 8

[99] Ruo Zhang, Ping-Sing Tsai, James Edwin Cryer, and Mubarak Shah. Shape from shading: A survey. *PAMI*, 21(8):690–706, 1999. 2

[100] Zhengyou Zhang. Flexible camera calibration by viewing a plane from unknown orientations. In *ICCV*, 1999. 3

[101] Zhiming Zhou, Guojun Chen, Yue Dong, David Wipf, Yong Yu, John Snyder, and Xin Tong. Sparse-as-possible svbrdf acquisition. *ACM Trans. on Graphics*, 35(6):189:1–189:12, 2016. 2

[102] Zhenglong Zhou and Ping Tan. Ring-light photometric stereo. In *ECCV*, 2010. 2

[103] Zhenglong Zhou, Zhe Wu, and Ping Tan. Multi-view photometric stereo with spatially varying isotropic materials. In *CVPR*, 2013. 2

[104] Michael Zollhöfer, Angela Dai, Matthias Innmann, Chenglei Wu, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Shading-based refinement on volumetric signed distance functions. *ACM Trans. on Graphics*, 34(4):96:1–96:14, 2015. 2

[105] Xinxin Zuo, Sen Wang, Jiangbin Zheng, and Ruigang Yang. Detailed surface geometry and albedo recovery from RGB-D video under natural illumination. In *ICCV*, 2017. 1, 2