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

Carolin Schmitt\textsuperscript{1,2,*} Simon Donn\textsuperscript{e1,2,*} Gernot Riegler\textsuperscript{3} Vladlen Koltun\textsuperscript{3} Andreas Geiger\textsuperscript{1,2}

\textsuperscript{1}Max Planck Institute for Intelligent Systems, Tübingen
\textsuperscript{2}University of Tübingen  \textsuperscript{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.

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 \cite{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 \cite{26, 40, 49} or inferred from images by assuming known \cite{15, 36, 65} or flat \cite{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.

\textsuperscript{*} Joint first author with equal contribution.

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 \cite{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 \cite{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.
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_svbdrf/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 \( \mathcal{I}_i : \mathbb{R}^2 \rightarrow \mathbb{R}^3 \) captured from \( N \) different views \( i \in \{1, \ldots, N\} \). 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 \( \mathbf{x}_p \in \mathbb{R}^3 \) 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 \( \pi_i : \mathbb{R}^3 \rightarrow \mathbb{R}^2 \).

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/index. Assuming a pinhole projection, the 3D location of surface point \( p \) is given by

\[
\mathbf{x}_p = \pi_i^{-1}(u_p, v_p, z_p) = \left( \frac{u_p - c_x}{f}, \frac{v_p - c_y}{f}, 1 \right) z_p
\]

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 \( p \), \( \pi_i^{-1} \) 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 \( f_p(\mathbf{n}_p, \omega^\text{in}, \omega^\text{out}) \) models the fraction of light that is reflected from incoming light direction \( \omega^\text{in} \) to outgoing light direction \( \omega^\text{out} \) given the surface normal \( \mathbf{n}_p \) at point \( p \). We use a modified version of the Cook-Torrance microfacet BRDF model [13]

\[
f_p(\mathbf{n}_p, \omega^\text{in}, \omega^\text{out}) = d_p + s_p \frac{D(r_p) G(\mathbf{n}_p, \omega^\text{in}, \omega^\text{out}, r_p)}{\pi(\mathbf{n}_p \cdot \omega^\text{in})(\mathbf{n}_p \cdot \omega^\text{out})}
\]

where \( D(\cdot) \) describes the microfacet slope distribution, \( G(\cdot) \) is the geometric attenuation factor, and \( d_p \in \mathbb{R}^3 \), \( s_p \in \mathbb{R}^3 \) and \( r_p \in \mathbb{R}^3 \) 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
albedo \( \mathbf{d}_p \) to vary freely per pixel \( p \), and model specular reflectance as a combination of \( T \) specular base materials

\[
\left( \mathbf{r}_p \right) = \sum_{t=1}^{T} \alpha^t_p \left( \mathbf{r}_t \right)
\]

with per-pixel BRDF weights \( \alpha^t_p \in [0, 1] \) and specular base materials \( \{ \{ \mathbf{s}_t, \mathbf{r}_t \} \}_{t=1}^{T} \). 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 \( \{ \{ \mathbf{s}_t, \mathbf{r}_t \} \}_{t=1}^{T} \) and \( \{ \mathbf{d}_p, \alpha^t_p \}_{p=1}^{P} \).

### 3.2. Model

This section describes our objective function. Let \( \mathcal{X} = \{ \{ (\mathbf{z}_p, \mathbf{n}_p, f_p) \}_{p=1}^{P}, \{ \mathbf{t}_i \}_{i=2}^{\infty} \} \) 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

\[
\mathcal{X}^* = \arg\min_{\mathcal{X}} \psi_{\mathcal{P}} + \psi_{\mathcal{G}} + \psi_{\mathcal{D}} + \psi_{N} + \psi_{M}
\]

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 \( \psi_{N} \) and material smoothness \( \psi_{M} \).

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

\[
\psi_p(\mathcal{X}) = \frac{1}{N} \sum_{i} \sum_{p} \left\| \varphi^i_p [\mathcal{I}_i(\mathbf{x}_p)] - \mathcal{R}_i(\mathbf{x}_p, \mathbf{n}_p, f_p) \right\|_1
\]

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

\[
\mathcal{R}_i(\mathbf{x}_p, \mathbf{n}_p, f_p) = \frac{f_p(\mathbf{n}_p, \omega^\text{in}_i(\mathbf{x}_p), \omega^\text{out}_i(\mathbf{x}_p)) \cdot \omega^\text{out}_i(\mathbf{x}_p)}{d_i(\mathbf{x}_p)^2} L
\]

where \( \omega^\text{in}_i(\mathbf{x}_p) \) denotes the direction of the ray from the surface point \( \mathbf{x}_p \) to the light source and \( \omega^\text{out}_i(\mathbf{x}_p) \) 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 \( \varphi^i_p \) in (5) disables occluded or shadowed observations i.e., we do not optimize for these regions. We set \( \varphi^i_p = 1 \) 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 \( \varphi^i_p = 0 \) 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 \( \{ n_p \} \) 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 \( \{ n_p \} \) and the cross product of the surface tangents at \( \{ x_p \} \):

\[
\psi_G(\mathcal{X}) = -\sum_p \tilde{n}_p^T \left( \frac{\partial z_p}{\partial x} \times \frac{\partial z_p}{\partial y} \right)_p \tag{7}
\]

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

\[
\frac{\partial z_p}{\partial x} \propto \left[ 1, 0, \nabla Z_1(\pi_1(\tilde{x}_p))^T [f/z_p, 0]^T \right]^T \tag{8}
\]

where \( \nabla Z_1(\pi_1(\tilde{x}_p)) \) 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 \( Z \) in the reference view \( i = 1 \) by regularizing our estimates \( z_p \) against it:

\[
\psi_D(\mathcal{X}) = \sum_p \| z_p - Z_1(u_p, v_p) \|_2^2 \tag{9}
\]

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 \( p \sim q \)

\[
\psi_G(\mathcal{X}) = \sum_{p \sim q} \| n_p - n_q \|_2^2 \tag{10}
\]

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

\[
\psi_M(\mathcal{X}) = \sum_p \left\| \alpha_p - \sum_q \alpha_q w_{q} k_{p,q} \right\|_1 - \sum_{p} \left\| \alpha_p - \frac{1}{P} \sum_{q} \alpha_q \right\|_1 \tag{11}
\]

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:

\[
k_{p,q} = \exp \left( -\frac{(x_p - x_q)^2}{2\sigma^2} - \frac{(d_p - d_q)^2}{2\sigma^2} \right) \tag{12}
\]

As the only informative regions are those that potentially observe a highlight, the weights \( w_q = \max_i \cos^{-1}(\mathbf{n}_q \cdot \mathbf{h}_q^i) \) 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 \( \mathbf{h}_q^i \) (i.e., the bisector between \( \omega^{in} \) and \( \omega^{out} \)) 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 \( Z = \{ z_p \} \) 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 \( Z \). 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.
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 \( s_t \) 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 \( x \), 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

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

space limitations, we provide more details about our system in the supplement.

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 \( \{x_p\} \) 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 \( n_p \) 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.
<table>
<thead>
<tr>
<th>Photometric Test Error</th>
<th>Overall</th>
<th>Specular</th>
<th>Non-Specular</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed Poses</td>
<td>1.210</td>
<td>3.349</td>
<td>1.151</td>
</tr>
<tr>
<td>Full Model</td>
<td>1.138</td>
<td>3.243</td>
<td>1.081</td>
</tr>
</tbody>
</table>

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 meaningful segmentations, as illustrated in Fig. 2, as well as more successful 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 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 \emph{disjoint} variant of our method that alternates between geometry and material.
updates. We also evaluate the improvement over the initial TSDF fusion using the implementation of Zeng et al. [98].

Our experimental evaluation 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.

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


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


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


