Multi-view 3D Reconstruction of a Texture-less Smooth Surface of Unknown Generic Reflectance

Ziang Cheng[1], Hongdong Li[1], Yuta Asano[2], Yinqiang Zheng[3], Imari Sato[2]
[1]Australian National University
[2]National Institute of Informatics,[3]The University of Tokyo, Japan
{ziang.cheng,hongdong.li}@anu.edu.au

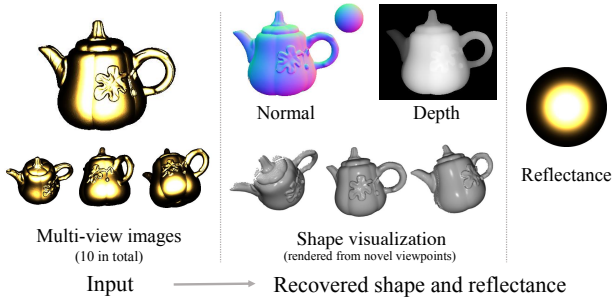


Figure 1: From a small set of input multi-view image (left), our method recovers the dense 3D object (middle) and its unknown generic surface reflectance (right).

Abstract

Recovering the 3D geometry of a purely texture-less object with generally unknown surface reflectance (e.g. non-Lambertian) is regarded as a challenging task in multi-view reconstruction. The major obstacle revolves around establishing cross-view correspondences where photometric constancy is violated. This paper proposes a simple and practical solution to overcome this challenge based on a co-located *camera-light scanner device. Unlike existing solutions, we do not explicitly solve for correspondence. Instead, we argue the problem is generally well-posed by multi-view geometrical and photometric constraints, and can be solved from a small number of input views. We formulate the reconstruction task as a joint energy minimization over the surface geometry and reflectance. Despite this energy is highly non-convex, we develop an optimization algorithm that robustly recovers globally optimal shape and reflectance even from a random initialization. Extensive experiments on both simulated and real data have validated our method, and possible future extensions are discussed.*

1. Introduction

3D reconstruction from multi-view images is one of the central problems in computer vision. Most traditional multi-view reconstruction methods such as SFM (structure from motion) often assume the scene or object to be reconstructed have distinctive features that are view-independent, so that cross-view feature correspondences can be readily established. However, this is not the case for many commonly-seen real-world objects or surfaces manifesting non-Lambertian reflectance. Traditional SFM methods are unable to reconstruct such texture-less surfaces with glossy appearance. The problem is even more challenging if the generic surface reflectance is unknown, in which case there is no apparent way to model how object's appearance changes with viewpoint.

Figure 2: Two experiment setups: Hardware used for capturing images under a co-located setup. A point light source is rigidly attached to camera lens with a small displacement.

By marrying photometric stereo with traditional multi-view methods, many papers have succeeded in overcoming parts of these challenges. Most of these methods are reliant on an external initialization of the 3D shape (e.g. [27, 6, 24]) to establish initial correspondences. Typically, finer-grained details are added incrementally to the recovered geometry. However, good initialization is not often guaranteed (*e.g.* many require initial shape from SFM pipelines, which are already vulnerable to textureless surface or specular highlights), and a large number of input images are often required. Additionally, many methods re-

sort to restrictive assumptions about the setup, objects and scenes. Common assumptions include *e.g.*, purely/almost Lambertian reflectance [15, 7, 8, 39, 3, 29, 19], planer object shape [9], known depth via an RGB-D sensor [31], or stereo vision with semi-static viewpoint but varying illumination [41, 13]. So far, direct multi-view reconstruction of textureless, glossy objects remains an open challenge that no method addresses well.

This paper presents a simple and practical solution to jointly recover high-fidelity surface geometry as well as unknown generic reflectance of a purely texture-less object. We advocate a co-located camera and light-source configuration, as shown in Fig. 2, where a point light source is rigidly and closely attached to the camera. Such scanner device is easily accessible with commodity hardware (*e.g.* a mobile phone camera with built-in flash).

Unlike existing photometric 3D reconstruction methods under fixed viewpoint, we allow the camera to move freely to leverage multi-view constraints. Our method is also different from existing multi-view methods mentioned above, in that we do not intend to establish explicit cross-view correspondences, either by feature matching or through shape initialization, since both are hard to obtain for a purely texture-less surface with arbitrary BRDF. Instead, we show that given a small number of views, shape and reflectance are already well-constrained by a physically-based image formation model. With this observation, we formulate reconstruction task as an energy minimization problem involving a single, unified objective. While this problem is still highly non-convex, we propose an effective optimization based approach that robustly reconstructs complex geometry as well as general reflectance without initial shape. Code and data will be available at https://github.com/za-cheng/PM-PMVS/.

2. Related work

Multi-view Photometric stereo. Multi-view photometric stereo (MVPS) methods often formulate the task as energy optimization, and solve it iteratively from a coarse initialization. The initialization is often obtained via SFM 3D reconstruction [36, 33] or from object's visual hull [12]. Many methods assume Lambertian reflectance, in which case surface normal can be solved under a linear system, to which specularities are simply discarded as outliers [16, 7, 39, 29, 19]. Recovered normal field is later used to refine the 3D shape geometry, and high-frequency shape details can be gradually recovered in an iterative manner. Under known global illumination, this can be further extended to analytical BRDF models, allowing recovery of reflectance as well [27]. Another notable branch of MVPS methods [41, 13] uses iso-depth constraints from a light ring to propagate initial sparse correspondences from SFM. When reflectance is Lambertian, surface details can also be reconstructed by fusing shape-from-shading with classical multiview stereo [11, 10, 22, 23]. However, for a textureless surface of specular material that extends outside viewing frustum, neither multi-view stereo nor visual hull is applicable, which puts a significant challenge on initializing existing MVPS methods. Additionally, MVPS methods generally require a large number of views (often a few hundreds), and most methods are designed for special scanner systems that are hard to build.

Co-located Photometric stereo. Similar to our configuration, Higo *et al*. [8] used controlled illumination consisting of a perspective camera with a rigidly attached point light source. They assume predominantly diffusive (i.e., Lambertian) reflectance to simplify surface normal solution, and treat specularities as outliers. With a similar co-located camera-light setup, Hui *et al*. [9] proposed a method for general spatially-varying BRDF (SVBRDF) sampling, but it is only applicable to planar surface where pixel correspondences can be easily found (via homography). Li *et al*. [14] addressed the same problem by using a single mobile phone image with the assistance of deep learning. In addition to its accessibility, a co-located setup can naturally reduce the 3D input space of isotropic BRDFs to a univariate one, which helps to improve estimation robustness even with less images. Nam *et al*. [24] later used a similar setup for joint reflectance and shape recovery. However, like most multi-view photometric methods, they require an initial shape input from a large set of images and cannot handle highly specular materials. Wang *et al*. [38] use a co-located light source to decouple normal and reflectance estimation in a standard photometric stereo setup. Schmitt and Donné *et al*. [31] used a handheld RGB-D method to improve camera pose estimation, where the depth sensor is used to provide shape initialization.

3. Problem Setup: Image formation

The overall setup of our camera and light-source is illustrated in Fig. 3, where a single point light source is rigidly attached to the camera with a small distance to the camera centre.

We assume the object's surface BRDF is uniform and isotropic, which under out setup reduces to a univariate function of incident/view angle (Fig. 4). We further assume the surface is smooth (or at least piece-wise smooth) so that surface normal vectors can be defined almost everywhere. While in this paper we mostly focus on solving spatially-uniform BRDF, our method (and its principle) can be extended to spatially-variant or SVBRDF as well. However, this is to be thoroughly addressed elsewhere to keep this paper concise and focused.

Consider multiple images of the surface K are given, each from a distinct viewpoint $m \in M$ with the above co-

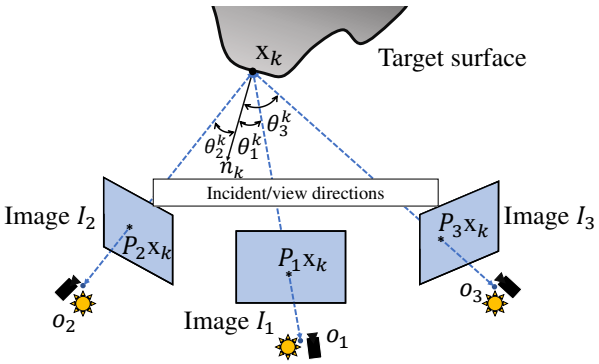


Figure 3: The geometric configuration of camera and light source in our reconstruction system. A light ray emitted from the point light o_i hits surface point k. The reflected ray reaches camera, also at o_i, through projection P_i forming multi-view observation from different viewpoints (i=1,2,...). Our task is to recover the surface

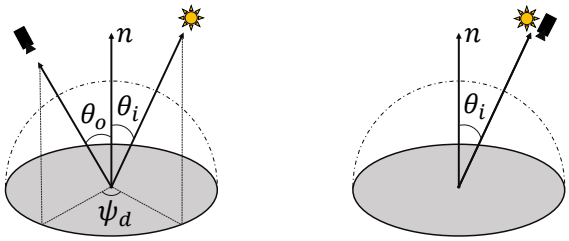


Figure 4: Left: Isotropic BRDF defined on three angular variables. Right: BRDF becomes univariate function under co-located setup.

located system. Denote $\mathbf{x}_k \in \mathbb{R}^3$ as the 3D world coordinates of the k-th surface point. Then, its radiance observed at view m is proportional to the corresponding raw pixel intensity in the m_{th} image I_m. This gives our photometric image formation equation:

$$\alpha_m^k I_m(P_m \mathbf{x}_k) = \alpha_m^k \frac{\gamma \rho(\theta_m^k)}{d(o_m, \mathbf{x}_k)^2}, \tag{1}$$

where $d(o_m, \mathbf{x}_k) = \|o_m - \mathbf{x}_k\|$ is the Euclidean distance between o_m and $\mathbf{x}_k$, and $\rho(\cdot)$ is the 1D BRDF[1], o_m is the centre of view m in world coordinates, and θ_m^k denotes the incident/view angle between vector $(\mathbf{x}_k - o_m)$ and surface normal $\mathbf{n}_k$. P_m is the camera perspective projection matrix. We assume all camera views $(P_m, m = 1...N)$ are geometrically calibrated, namely they are known parameters. This is easy to achieve in practice, by using any off-the-shelf camera calibration software toolbox(*e.g.* [20]). γ is a constant factor proportional to the brightness of light source and the response function of the camera, and α_m^k encodes the visibility of surface point k in view m (*i.e.*, $\alpha_m^k = 1$ if k is visible in view m and $\alpha_m^k = 0$ if otherwise,*e.g.* caused by occlusion or image boundary cropping).

To address the high dynamic range especially when imaging a non-Lambertian (*e.g.* highly specular) surface, the above photometric image formation equation is often rewritten in the log-space, namely,

$$\alpha_m^k \big(\log \rho(\theta_m^k) - \log I_m(P_m \mathbf{x}_k) - \log \|o_m - \mathbf{x}_k\|_2^2 + \log \gamma \big) = 0. \tag{2}$$

This log-scale form helps to improve numeric condition of the problem (ref. Nielsen *et al.* [25] for more details).

4. Energy Minimization: Joint shape and BRDF recovery

Recall our goal is to recover the unknown BRDF and 3D shape from multi-view observations. We use a coefficient vector $\mathbf{c}$ to parameterize a BRDF function (*c.f.* Sec. 4.1). The shape is represented as a set of points K indexed by pixels in the reference frame[2]. The surface shape is thus modelled as depth and normal maps in the reference frame, *i.e.* z_k and n_k for all $k \in K$. Under perspective camera model, the relation between world coordinates $\mathbf{x}_k$ and depth z_k can be readily established as $\mathbf{x}_k = z_k(P_{\text{ref}}^+ p_k - o_1) + o_1$ where P_{ref}^+ is the inverse projection that maps reference frame pixel p_k in image coordinates onto the unit depth plane in world coordinates.

Our energy function is defined as a weighted sum of three energy terms: photometric term E_p, shape term E_s and BRDF term E_c.

$$E(\mathbf{n}, \mathbf{z}, \mathbf{c}) = E_p(\mathbf{n}, \mathbf{z}, \mathbf{c}) + \lambda_s E_s(\mathbf{n}, \mathbf{z}) + \lambda_c E_c(\mathbf{c}). \tag{3}$$

The first term represent the photometric constraint Eq. (2), and the latter two can be viewed as regularizers on shape and reflectance respectively.

4.1. BRDF parameterization (E_c)

Before we introduce our energy model, let us first define our non-parametric BRDF representation. Recent work on BRDF parameterization [25, 40] suggested that a wide range of real-world BRDFs can be approximated by a linear combination of a compact set of BRDF bases with high accuracy in the log space. Specifically, $\log \rho(\cdot)$ is approximated by

$$\log \rho(\cdot) \approx \mathtt{D}(\cdot)\mathbf{c} + \mu(\cdot) \tag{4}$$

where $\mathbf{c} \in \mathbb{R}^N$ is the linear mixing coefficients, and μ is the average log-BRDF. $\mathtt{D} = [d_1, d_2, .., d_N]$ is a set of pre-learned BRDF basis functions.

Given a collection of real BRDFs (*e.g.* MERL [21]), one can learn the bases as the leading N eigenfunctions. We further weight basis d_i by its square-rooted eigenvalues to improve conditioning, as suggested by [25]. Overall the BRDF term becomes:

$$E_c(\mathbf{c}) = \|\log \rho - \mathtt{D}\mathbf{c} - \mu\|^2 + \|\mathbf{c}\|^2 \approx \|\mathbf{c}\|^2. \tag{5}$$

[1] We assume the cosine fall-off factor is subsumed in $\rho()$.

[2] without loss of generality, the first input image is used as the reference

This term effectively encourages $\mathbf{c}$ to follow a spherical Gaussian prior in our non-parametric BRDF space. A similar Tikhonov regularizer has been used to reduce the BRDF sampling as well [40].

4.2. Photometric term E_p

From image formation model Eq. (2), a multi-view 'rendering loss' of point cloud K could be straightforwardly derived as

$$\mathcal{E}_{render} = \frac{1}{|K||M|}\sum_{k,m}\alpha_m^k L_\delta\big(\Phi_m(n_k, z_k, \mathbf{c})\big), \quad (6)$$

where $L_\delta(\cdot)$ denotes the robust Huber loss (with clipping parameter δ), $\alpha_m^k \in \{0,1\}$ is the visibility of point k in view m and $\Phi_m(n_k, z_k, c)$ measures the log-scale difference between expected scene radiance and true pixel intensities

$$\Phi_m(n_k, z_k, \mathbf{c}) = \mathtt{D}(\theta_m^k)\mathbf{c} + \mu(\theta_m^k) - \log\Big(\frac{1}{\gamma}I_m(P_m\mathbf{x}_k)\|o_m - \mathbf{x}_k\|_2^2\Big). \quad (7)$$

A major challenge for computing Eq.(6) revolves around the visibility mask α_m^k, which depends not only on viewpoint but also the unknown shape to be solved for, hence cannot be computed a priori. To overcome this difficulty, we use a simple heuristic to select for each surface point k a subset of $\mathcal{M}$ images with the smallest photometric errors. We assume every surface points k should be visible in at least $\mathcal{M}$ out of M views (*i.e.* $\forall k \in K \ \sum_m \alpha_m^k \geq \mathcal{M}$), where $\mathcal{M} < |M|$ is a pre-set constant. Mathematically, we define our photometric energy term as

$$E_p(\mathbf{n}, \mathbf{z}, \mathbf{c}) = \frac{1}{|K|\mathcal{M}}\sum_{k\in K}\min_{|M_k|=\mathcal{M}}\sum_{m\in M_k} L_\delta\big(\Phi_m(n_k, z_k, \mathbf{c})\big). \quad (8)$$

Notably, the min-sum selector M_k, in its way of handling occluded or out-of-view points, is analogue to least trimmed squares — an objective function commonly seen in robust regression.

4.3. Shape term E_s

A prerequisite for image formation model is the surface should be smooth, hence its normal can be defined almost everywhere and is perpendicular to tangent vectors. We design E_s following this observation, to encourage surface smoothness and local integrity:

$$E_s(\mathbf{n}, \mathbf{z}) = \frac{1}{|K||\mathcal{N}_k|}\sum_k\sum_{j\in\mathcal{N}_k}\big(n_k^T(\mathbf{x}_k - \mathbf{x}_j)\big)^2, \quad (9)$$

where $\mathcal{N}_k$ is the 4-neighbor of pixel k in the reference frame.

5. Solving the energy minimization

The above energy function is highly non-convex due to the view-dependent non-Lambertian BRDF and the arbitrary image measurements in the scene. Previous photometric methods tackled this non-convexity either by using overly simplistic reflectance model (*e.g.* pure Lambertian) or by assuming the availability of high quality initialization, if not both.

In this paper, we do not rely on above assumptions. Our energy minimization algorithm is based on *coordinate descent*, alternating between two sub-problems of solving BRDF and solving shape parameters respectively (see Sec. 5.1 and Sec. 5.2). Therefore it can be initialized from an inexact estimation of either shape or BRDF. In fact, we show it can converge quickly and correctly even from a *null* initialization, namely, starting from the triviality of $\mathbf{c} = 0$ with no initial shape.

5.1. Solve for BRDF, fixing Shape

Suppose a current 3D shape estimation is given and fixed, we solve for the BRDF by minimizing Eq. (3), *i.e.*,

$$\min_{\mathbf{c},\gamma} E_p(\mathbf{n}, \mathbf{z}, \mathbf{c}) + \lambda_c E_c(\mathbf{c}). \quad (10)$$

Here we keep the minimum set M_k constant during the optimization, in which case $E_p(\mathbf{n}, \mathbf{z}, \cdot)$ becomes convex and can be globally minimized. We note that in practice such approximation has minimal impact on the estimation due to problem being well-constrained on $\mathbf{c}$, and the approximation is indeed an upper bound of true energy. We employ a standard L-BFGS optimizer [26] and solve the camera response rate γ together with $\mathbf{c}$.

5.2. Solve for Shape, fixing BRDF

With fixed BRDF, surface shape is solved by minimizing Eq. (3) w.r.t. $\mathbf{n}, \mathbf{z}$. This is a highly challenging minimization problem due to the non-convexity of the image formation function $I_m(P_m\cdot)$ and BRDF $\rho_c(\cdot)$. In our experiments we found that conventional optimization algorithms (*e.g.* gradient-descent or quasi-Newton) almost always fail unless provided with a high quality initialization.

Formally, we seek to solve the following minimization problem:

$$\min_{\mathbf{n},\mathbf{z}} E_p(\mathbf{n}, \mathbf{z}, \mathbf{c}) + \lambda_s E_s(\mathbf{n}, \mathbf{z}). \quad (11)$$

To do this, we introduce a set of auxiliary variables $\tilde{\mathbf{z}} = [\tilde{z}_1, ..., \tilde{z}_K]^T$ to decouple E_p and E_s, and employ a quadratic penalty method (QPM) [35] to relax the 'hard' constraint $\tilde{\mathbf{z}} - \mathbf{z} = 0$

$$E_{\text{QPM}}(\mathbf{n}, \mathbf{z}, \tilde{\mathbf{z}}) = E_p(\mathbf{n}, \mathbf{z}, \mathbf{c}) + \lambda_s E_s(\mathbf{n}, \tilde{\mathbf{z}}) + \sigma^{(i)}\|\tilde{\mathbf{z}} - \mathbf{z}\|^2, \quad (12)$$

where $\sigma^{(i)} = \kappa\sigma^{(i-1)}$ is a penalty coefficient that increases exponentially w.r.t. i by some factor $\kappa > 1$. As i grows, violations of $\tilde{\mathbf{z}} - \mathbf{z} = 0$ are penalized with increased severity until the constraint is satisfied. In this paper we use fixed $\kappa = 1.3$.

The purpose for above relaxation is that unlike Eq.(11), Eq.(12) is now convex (quadratic) w.r.t. $\tilde{\mathbf{z}}$, and *k-separable* w.r.t. $\mathbf{n}, \mathbf{z}$. In other words, the objective w.r.t. $\mathbf{n}, \mathbf{z}$ can be re-written as a summation over K mutually independent sub-energy terms [3]

$$E_{\text{QPM}}(\mathbf{n}, \mathbf{z}, \tilde{\mathbf{z}}) = \sum_{k \in K} E^k_{\text{QPM}}(\mathbf{n}_k, z_k, \tilde{\mathbf{z}}), \tag{13}$$

where each term $E^k_{\text{QPM}}(\cdot, \cdot, \tilde{\mathbf{z}})$ has a small input space that can be directly searched. Therefore $E_{\text{QPM}}(\cdot, \cdot, \tilde{\mathbf{z}})$ as a whole can be globally minimized by *e.g.* exhaustive search within complexity $O(K)$[4]. This is a crucial step as it allows one to overcome the non-convexity of original energy (most of which is in $\mathbf{n}, \mathbf{z}$ dimensions) in a tractable manner.

We solve the QPM optimization by alternating coordinate descent between $\{\mathbf{n}, \mathbf{z}\}$ and $\tilde{\mathbf{z}}$, as listed in Alg. 1. Similar to the arguments made above, the optimal $\tilde{\mathbf{z}}$ is amenable to closed-form solution by the sparse least square LSQR algorithm [28]. Conversely, optimization of $\{\mathbf{n}, \mathbf{z}\}$ is non-convex, but can be decomposed into K independent sub-problems.

Here we solve the latter using a randomized search approach inspired by PatchMatch [1, 2], which iterates between two steps: *propagation* and *randomized search*. During *propagation* step, for every k we attempt to improve E^k_{QPM} using its neighbors' normal and depth $\{\mathbf{n}_j, z_j | j \in \mathcal{N}_k\}$ as candidates. During the *randomized search*, we attempt to improve E^k_{QPM} with random candidates of $\mathbf{n}_k, z_k$. In both steps, the best candidate is kept and carried to the next iteration. In practice, we found PatchMatch to be significantly more efficient than exhaustive search, thanks to the spatial smoothness of shape variables. An optimal solution can often be found in just 10 to 15 iterations. Note that PatchMatch does not need an initialization. Rather it starts from a random initial point. We refer the readers to Barnes *et al*. [1, 2] for details about PatchMatch.

An overview of relaxed minimization is presented in Alg. 1. Note the solution at least converges to a local minimum, as eventually neither PatchMatch nor LSQR increases energy when σ no longer changes. In practice, we found the solution turned out to be almost always globally optimal.

[3] See supplemental for proof.

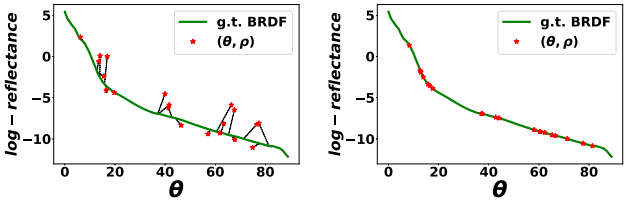
[4] Global optimality is guaranteed under the weak assumption that the depth z_k is bounded and energy is bounded Lipschitz continuous, see supplemental for proof.

Algorithm 1 Solution for Eq. (11) by solving a series of quadratic coupling sub-problems.

function SOLVESHAPE($\mathbf{n}, \mathbf{z}, \sigma^{(0)}$)
  $\tilde{\mathbf{z}} = \mathbf{0}$
  $\sigma = \sigma^{(0)}$ ▷ initially very small
  repeat
    $\mathbf{n}, \mathbf{z}$ = PATCHMATCH(min_func=$E_{\text{QPM}}(\cdot, \cdot, \tilde{\mathbf{z}})$)
    $\tilde{\mathbf{z}}$ = LSQR(min_func=$E_{\text{QPM}}(\mathbf{n}, \mathbf{z}, \cdot)$)
    if $\|\mathbf{z} - \tilde{\mathbf{z}}\|$ is not sufficiently small **then**
      $\sigma = \kappa\sigma$
      continue
    end if
  until Energy converges
  return $\mathbf{n}, \mathbf{z}$
end function

5.3. Analysis

Before presenting our experiment results,let us analyze the existence of optimal solution. Specifically, we will show that ground-truth (shape and BRDF) is indeed a valid optimizer for the energy function, and conversely, minimizing this energy to its optimum will lead to the true solution. We do so by considering two reflectance estimations: one is the ground-truth BRDF itself, and the other is inferred from the estimated shape $\{z_k, \mathbf{n}_k\}$ by solving Eq. (2), as illustrated in Fig. 5. The equality of Eq. (2) states that the two estimations must agree with each other. Otherwise a discrepancy generally exists between the two, which contributes to a non-minimum energy value. By minimizing the energy, such discrepancy is also reduced, eventually leading to the true BRDF and true shape. (The reader is also referred to the curves obtained in our real experiments in Fig .9 for a better visualization.)

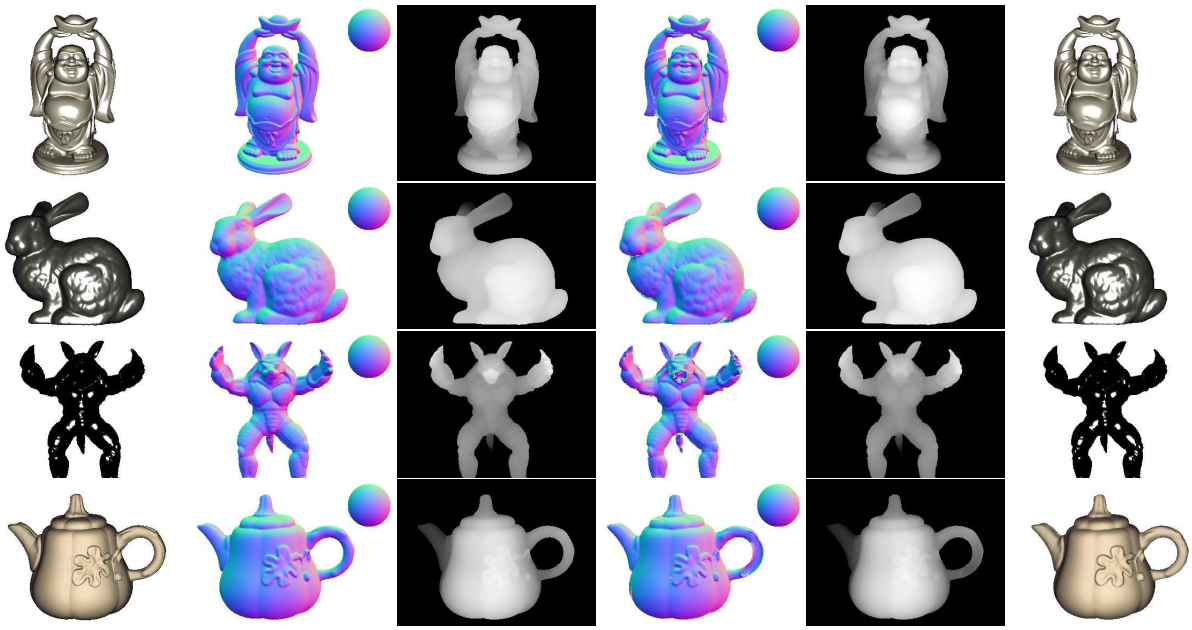


(a) with wrong $\mathbf{x}_k, n_k$ (b) with ground-truth shape

Figure 5: Shape and BRDF visualized in θ-reflectance domain. The green curve is the ground truth BRDF and red points are inferred from shape under Eq. (2). The latter is nonconforming to any BRDF if shape is given wrong values (5a). On the other hand, the ground-truth shape corresponds to a true minimizer as shown in (5b).

6. Experiments

To validate the proposed method, we conducted experiments on both synthetic and real images of multi-view photometric observations.



(a) ref. view (b) g.t. normal (c) g.t. depth (d) est. normal (e) est. depth (f) re-render

Figure 6: Results on Synthetic objects. See supplemental for more visual results.

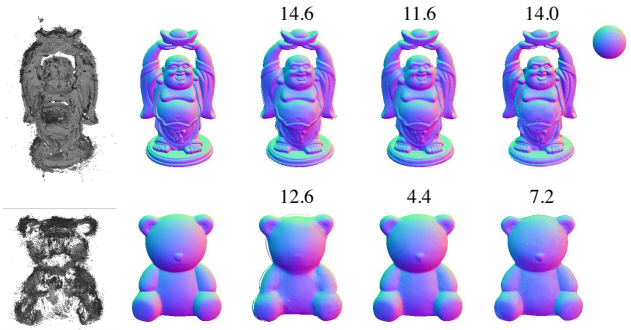


Figure 7: Comparison on Diligent-MV dataset [13] with normal errors listed on top. Note our method received far less images than the other two methods, and does not require high quality initial shape.

6.1. Synthetic experiments

For synthetic experiments, our aim is to quantitatively verify the effectiveness of our method under different imaging conditions. Throughout our experiments, we use the same and fixed set of parameter settings, *i.e.*, $N = 15$, $\delta = 0.1$, $\lambda_s = 10^6$, $\lambda_c = 0.005$, suggesting the method is rather agnostic or robust to meta-parameters. Our energy minimization algorithm was always initialized from null BRDF *i.e.* $\mathbf{c} = \mathbf{0}$ and without initial shape.

We render multiple 3D mesh models with BRDFs sampled form the MERL database [21]. To validate the BRDF models, we perform multiple rounds of cross validation where 95 out of the 100 materials in MERL were used to learn the bases (or dictionary) $\mathtt{D}$ and the remaining five BRDFs for rendering and testing. We position the virtual camera approximately one world unit (metre) away from mesh in reference frame, and all objects (except the open surface) are scaled so that they span 0.25 unit length (25 centimetres) in whichever is the greatest of its X, Y and Z dimensions. We render only 10 views of the target objects with co-located point light source.

Some reconstructions obtained from our method are visualized in Fig 6, where we test the performance on four models 'buddha, bunny, armadillo, teapot' from Stanford dataset [4] and Diligent dataset [34]. Additionally, we render an open and smooth surface '*himmelblau*' (refer to supplemental for details). In table-1 we list quantitative performance w.r.t. ground truth shape and BRDF. We measure the mean errors for recovered surface normal in degrees, depth map (in world units), and log-BRDF (absolute difference averaged over input angles in $[0, \pi/2)$). Note the mean normal and depth errors are mostly caused by heavy occlusion on objects' boundaries where the shape cannot be exactly recovered, while the median errors are mush smaller.

6.2. Comparison

To the best of our knowledge, our method is original and few previous paper had attempted at this challenging task in the same setting. This makes a direct and fair comparison difficult. However, to provide the reader a sense of the performance of our 3D reconstruction, we offer comparison with two state-of-the-art methods – Park *et al*. [29] and Li *et al*. [13] – on Diligent-MV dataset [13]. We note that this re-

quires us to modify our method for a detachable light source (*i.e.* a non-co-located setup), and model a higher dimensional BRDF. In this experiment, we follow the bi-variate BRDF approximation, and incorporate a second difference angle in our BRDF formulation [30].

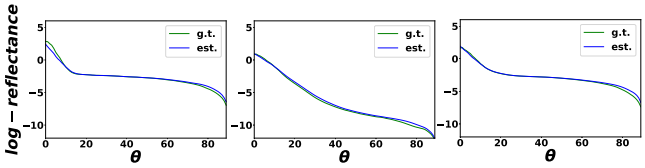
Furthermore, we note following differences between our methods and [29, 13]: (1) our method is based on 21 input images from 3 viewpoints, while [29, 13] received 1920 images from 20 viewpoints (2) we reconstruct an oriented point cloud indexed by reference pixels, while [29, 13] outputs a water tight triangle mesh (3) our method is randomly initialized, while [29, 13] received high quality initial shape from MVS pipeline [5] with human correspondence labeling involved.

Fig 7 illustrates the reconstructed normal maps on two real world models *Buddha* and *Bear*, with corresponding mean normal error listed on top. We also include the results from COLMAP [32] as an SFM baseline. We are able to achieve comparable performance to [29, 13] despite using far less images and unaided by initial geometry. Compared to COLMAP [32], we reconstruct a dense point cloud of arguably better quality.

Table 1: Error metrics on different target objects. Note we erode foreground region by 2 pixels for evaluation purpose since object's boundaries are often heavily self-occluded.

Models	Normal in degrees		Depth × 1000		BRDF × 10	
	median	mean	median	mean	median	mean
Buddha	1.87	5.59	0.45	1.36	0.37	1.44
Bunny	1.33	5.27	0.42	0.90	0.30	2.81
Armadillo	1.40	7.81	0.35	1.57	0.66	1.33
Teapot	0.67	2.64	0.51	0.82	0.38	0.55
Himmelblau	1.45	2.06	1.28	5.40	0.67	1.94
Overall	1.36	5.65	0.59	1.40	0.69	1.56

We also present examples of recovered (non-Lambertian) BRDFs compared with their corresponding ground-truths, as shown in Fig-8.



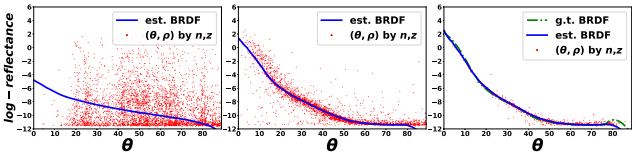
(a) alumina-oxide (b) black-phenolic (c) white-marble

Figure 8: Recovered BRDF curves versus the ground truth BRDFs for three different materials.

6.2.1 Convergence Analysis

As our energy model and minimization algorithm are mostly heuristic, it is difficult to prove the convergence on shape and BRDF metrics compared to ground truths, and such proof is beyond the scope of this paper. Instead, we offer to verify the convergence experimentally.

Fig. 9 depicts how well the image formation equation (Eq. (2)) is satisfied as the energy decreases. The vertical distances between each point and the predicted BRDF curve contribute to the overall inequality of multi-view photometric constraint Eq. (2). As iterations increase, this distance gradually decreases and the energy is minimized. The final solution in Fig. 9c is well-constrained and close to the true BRDF. Fig. 10 illustrates error metrics indeed converge as a function of iterations.



(a) 1_{st} iteration (b) 10_{th} iteration (c) 50_{th} iteration

Figure 9: Visualization of how multi-view photometric constraints are gradually approached as the algorithm iterates. Red points are the reflectances retrieved from current shape, and the blue curve is the fitted BRDF curve. Both fittings gradually converge to ground-truth.

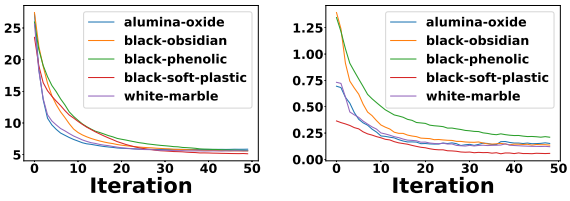
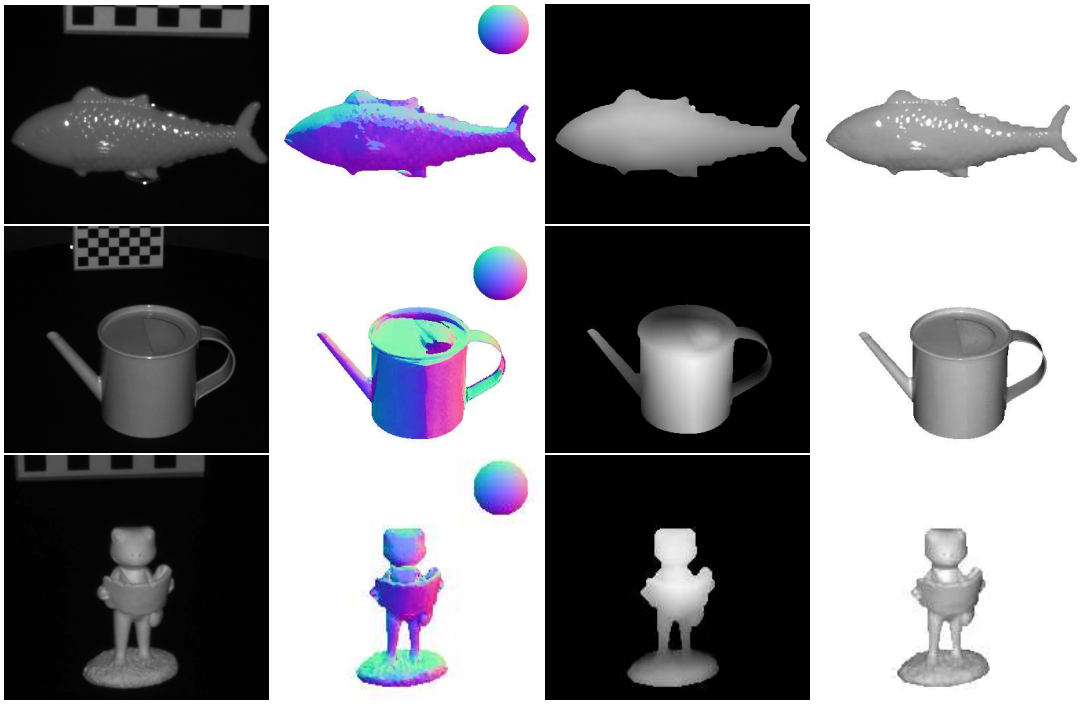


Figure 10: Left: normal angular error and Right: log-BRDF error (averaged over input angles) versus iterations.

6.3. Tests on real images

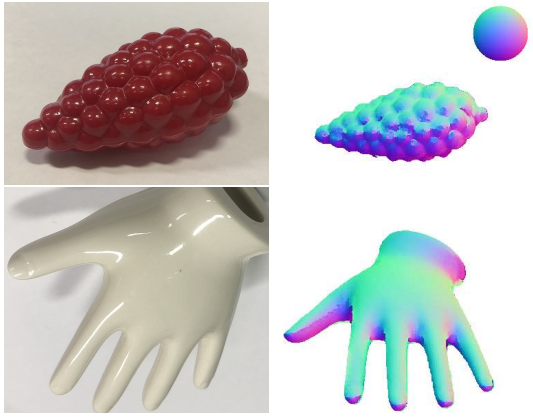
In this subsection we qualitatively evaluate the performance of proposed algorithm on real-world images. To build the co-located camera and light source, we rigidly attached a universal light source onto the camera lens, as shown in Fig. 2. Similar to synthetic experiments, we take 20 images at varying camera positions in front of each target object. Since our method relies on photometric measurements, throughout our experiments, we use RAW image format and ensure that the measured image intensity is produced from a linear response function. This is readily accessible for commodity DSLR cameras and many smartphone cameras.

For obtaining extrinsic camera parameters, a standard checkerboard is placed near the object, and intrinsic and extrinsic parameters are solved by minimizing the reprojection error on corner points (cf. Fig. 11) [17, 18, 37]. Input



(a) ref.view (b) est. normal (c) est. depth (d) re-rendered

Figure 11: Input real image, recovered normal map, depth map, and re-rendered shape reconstruction.



(a) real shape (b) est. normal

Figure 12: Naturally lit real world objects versus reconstructed shapes.

images are then rectified and foreground regions are manually cropped out. More experimental results on real data are shown in Fig. 11 and 12. It is clear that our method achieves fine-grained reconstruction; even small details (such as the tiny surface bumps) are recovered vividly.

7. Conclusion

This paper has presented a new multi-view photometric 3D reconstruction method built upon a simple and practical camera-light configuration. It is able to recover fine-detailed 3D shape of a purely texture less surface with unknown arbitrary (non-Lambertian) reflectances, from a small set of multi-view input images. Our key contribution is a new optimization procedure that solves the challenging (highly non-convex) energy minimization task effective and optimally, without proper initialization. Our method obtains visually compelling results on both synthetic data and real images. Possible future extensions include to relax the assumptions about the scene, such as isotropic BRDF, uniform material, or point light source. The co-located configuration may also be relaxed. One limitation of our methods is that it can only handle open smooth (or piecewise smooth) surface visible by the reference camera view, and assume uniform reflectance on its surface. While we are working on relaxing these limitations, we hope this work may inspire future researchers working in the field.

8. Acknowledgement

This research is funded in part by the ARC Centre of Excellence for Robotics Vision (CE140100016), ARC-Discovery (DP 190102261), JSPS KAKENHI (JP20H05951) and by the Ministry of Education, Science, Sports and Culture Grant-in-Aid for Scientific Research on Innovative Areas (JP15H05918). We would like to thank Liu Liu for his support in camera calibration.

References

[1] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.*, 28(3):24, 2009. 5

[2] Connelly Barnes, Eli Shechtman, Dan B Goldman, and Adam Finkelstein. The generalized patchmatch correspondence algorithm. In *European Conference on Computer Vision*, pages 29–43. Springer, 2010. 5

[3] Mate Beljan, Jens Ackermann, and Michael Goesele. Consensus multi-view photometric stereo. In *Joint DAGM (German Association for Pattern Recognition) and OAGM Symposium*, pages 287–296. Springer, 2012. 2

[4] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *SIGGRAPH '96*, 1996. 6

[5] Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 873–881, 2015. 7

[6] S. Georgoulis, M. Proesmans, and L. V. Gool. Tackling shapes and brdfs head-on. In *2014 2nd International Conference on 3D Vision*, volume 1, pages 267–274, 2014. 1

[7] Carlos Hernandez Esteban, George Vogiatzis, and Roberto Cipolla. Multiview photometric stereo. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(3):548–554, Mar. 2008. 2

[8] T. Higo, Y. Matsushita, N. Joshi, and K. Ikeuchi. A hand-held photometric stereo camera for 3-d modeling. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1234–1241, Sep. 2009. 2

[9] Zhuo Hui, Kalyan Sunkavalli, Joon-Young Lee, Sunil Hadap, Jian Wang, and Aswin C Sankaranarayanan. Reflectance capture using univariate sampling of brdfs. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5362–5370, 2017. 2

[10] Kichang Kim, Akihiko Torii, and Masatoshi Okutomi. Multi-view inverse rendering under arbitrary illumination and albedo. In *European conference on computer vision*, pages 750–767. Springer, 2016. 2

[11] Fabian Langguth, Kalyan Sunkavalli, Sunil Hadap, and Michael Goesele. Shading-aware multi-view stereo. In *European Conference on Computer Vision*, pages 469–485. Springer, 2016. 2

[12] Aldo Laurentini. The visual hull concept for silhouette-based image understanding. *IEEE Transactions on pattern analysis and machine intelligence*, 16(2):150–162, 1994. 2

[13] M. Li, Z. Zhou, Z. Wu, B. Shi, C. Diao, and P. Tan. Multi-view photometric stereo: A robust solution and benchmark dataset for spatially varying isotropic materials. *IEEE Transactions on Image Processing*, 29:4159–4173, 2020. 2, 6, 7

[14] Zhengqin Li, Kalyan Sunkavalli, and Manmohan Chandraker. Materials for masses: Svbrdf acquisition with a single mobile phone image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 72–87, 2018. 2

[15] Li Zhang, Curless, Hertzmann, and Seitz. Shape and motion under varying illumination: unifying structure from motion, photometric stereo, and multiview stereo. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 618–625 vol.1, Oct 2003. 2

[16] Jongwoo Lim, Jeffrey Ho, Ming-Hsuan Yang, and David Kriegman. Passive photometric stereo from motion. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pages 1635–1642. IEEE, 2005. 2

[17] Liu Liu, Hongdong Li, and Yuchao Dai. Efficient global 2d-3d matching for camera localization in a large-scale 3d map. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2372–2381, 2017. 7

[18] Liu Liu, Hongdong Li, Yuchao Dai, and Quan Pan. Robust and efficient relative pose with a multi-camera system for autonomous driving in highly dynamic environments. *IEEE Transactions on Intelligent Transportation Systems*, 19(8):2432–2444, 2017. 7

[19] Fotios Logothetis, Roberto Mecca, and Roberto Cipolla. A differential volumetric approach to multi-view photometric stereo. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1052–1061, 2019. 2

[20] Matlab optimization toolbox, 2020. The MathWorks, Natick, MA, USA. 3

[21] Wojciech Matusik, Hanspeter Pfister, Matt Brand, and Leonard McMillan. A data-driven reflectance model. *ACM Transactions on Graphics*, 22(3):759–769, July 2003. 3, 6

[22] Daniel Maurer, Yong Chul Ju, Michael Breuß, and Andrés Bruhn. Combining shape from shading and stereo: A joint variational method for estimating depth, illumination and albedo. *International Journal of Computer Vision*, 126(12):1342–1366, 2018. 2

[23] Jean Mélou, Yvain Quéau, Fabien Castan, and Jean-Denis Durou. A splitting-based algorithm for multi-view stereopsis of textureless objects. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 51–63. Springer, 2019. 2

[24] Giljoo Nam, Joo Ho Lee, Diego Gutierrez, and Min H. Kim. Practical svbrdf acquisition of 3d objects with unstructured flash photography. *ACM Trans. Graph.*, 37(6), Dec. 2018. 1, 2

[25] Jannik Boll Nielsen, Henrik Wann Jensen, and Ravi Ramamoorthi. On optimal, minimal brdf sampling for reflectance acquisition. *ACM Transactions on Graphics (TOG)*, 34(6):186, 2015. 3

[26] Jorge Nocedal. Updating quasi-newton matrices with limited storage. *Mathematics of computation*, 35(151):773–782, 1980. 4

[27] Geoffrey Oxholm and Ko Nishino. Multiview shape and reflectance from natural illumination. volume 7572, pages 528–541, 10 2012. 1, 2

[28] Christopher C. Paige and Michael A. Saunders. Lsqr: An algorithm for sparse linear equations and sparse least squares. *ACM Trans. Math. Softw.*, 8(1):43–71, Mar. 1982. 5

[29] Jaesik Park, Sudipta N. Sinha, Yasuyuki Matsushita, Yu-Wing Tai, and In So Kweon. Robust multiview photometric stereo using planar mesh parameterization. *IEEE Transactions of Pattern Analysis and Machine Intelligence (TPAMI)*, 2016. 2, 6, 7

[30] Szymon M Rusinkiewicz. A new change of variables for efficient brdf representation. In *Eurographics Workshop on Rendering Techniques*, pages 11–22. Springer, 1998. 7

[31] Carolin Schmitt, Simon Donne, Gernot Riegler, Vladlen Koltun, and Andreas Geiger. On joint estimation of pose, geometry and svbrdf from a handheld scanner. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[32] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 6, 7

[33] Steven M Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 519–528. IEEE, 2006. 2

[34] B. Shi, Z. Mo, Z. Wu, D. Duan, S. Yeung, and P. Tan. A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):271–284, Feb 2019. 6

[35] Frank Steinbrücker, Thomas Pock, and Daniel Cremers. Large displacement optical flow computation withoutwarping. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1609–1614. IEEE, 2009. 4

[36] Shimon Ullman. The interpretation of structure from motion. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 203(1153):405–426, 1979. 2

[37] Dong Wang, Quan Pan, Chunhui Zhao, Jinwen Hu, Liu Liu, and Limin Tian. Slam-based cooperative calibration for optical sensors array with gps/imu aided. In *2016 International conference on unmanned aircraft systems (ICUAS)*, pages 615–623. IEEE, 2016. 7

[38] Xi Wang, Zhenxiong Jian, and Mingjun Ren. Non-lambertian photometric stereo network based on inverse reflectance model with collocated light. *IEEE Transactions on Image Processing*, 29:6032–6042, 2020. 2

[39] Chenglei Wu, Yebin Liu, Qionghai Dai, and Bennett Wilburn. Fusing multiview and photometric stereo for 3d reconstruction under uncalibrated illumination. *IEEE transactions on visualization and computer graphics*, 17:1082–95, 08 2011. 2

[40] Zexiang Xu, Jannik Boll Nielsen, Jiyang Yu, Henrik Wann Jensen, and Ravi Ramamoorthi. Minimal brdf sampling for two-shot near-field reflectance acquisition. *ACM Transactions on Graphics (TOG)*, 35(6):188, 2016. 3, 4

[41] Z. Zhou, Z. Wu, and P. Tan. Multi-view photometric stereo with spatially varying isotropic materials. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1482–1489, June 2013. 2