Shape from Shading for Robotic Manipulation

Arkadeep Narayan Chaudhury Leonid Keselman
Christopher G. Atkeson
The Robotics Institute, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, Pennsylvania 15213, USA
{arkadeepnc, leonidk, cga}@cmu.edu

Abstract

Controlling illumination can generate high quality information about object surface normals and depth discontinuities at a low computational cost. In this work we demonstrate a robot workspace-scaled controlled illumination approach that generates high quality information for table top scale objects for robotic manipulation. With our low angle of incidence directional illumination approach, we can precisely capture surface normals and depth discontinuities of monochromatic Lambertian objects. We show that this approach to shape estimation is 1) valuable for general purpose grasping with a single point vacuum gripper, 2) can measure the deformation of known objects, and 3) can estimate pose of known objects and track unknown objects in the robot's workspace.

1. Introduction

Imagine the following task: you are trying to find a tiny screw on the ground, but the color contrast between the screw and the ground is poor, or the floor has a complex texture. You might turn to a flashlight to aid you in this task, illuminating part of the floor with a low angle of incidence, looking for shadow cues as you hunt for the screw.

Now consider the task of examining the quality of examining the quality of stucco [48] or putty applied to a surface; here, a low angle of incidence light (as in the first task) can highlight high spatial frequency surface discontinuities such as surface roughness and features like creases. A directional source of light (e.g. sunlight throughout the day), meanwhile, can highlight low spatial frequency information such as the curvature of the wall.

This work is motivated by the observation that changing the direction of illumination can often highlight object surface features and depth discontinuities on the object surface and between the object and its surroundings. These surface features often lead to shadows and illuminated patches when viewed with a camera. Shading and shadows along with an illumination model can aid in reasoning about the surface properties of an object and help us decide how to interact with it. Although classical techniques in multi-view geometry [41], shape from texture [3,45], high performance area scanners [1,2] and tactile sensing [18,55] exist for solving similar problems, multi-modal sensing adds complexity to the problem by requiring the camera poses and object features to be tracked across views and sensors. Coordinating tactile sensing with vision [14] adds the additional complexities of discovering correspondences between cameras and tactile sensor data and accounting for tactile sensor drift.

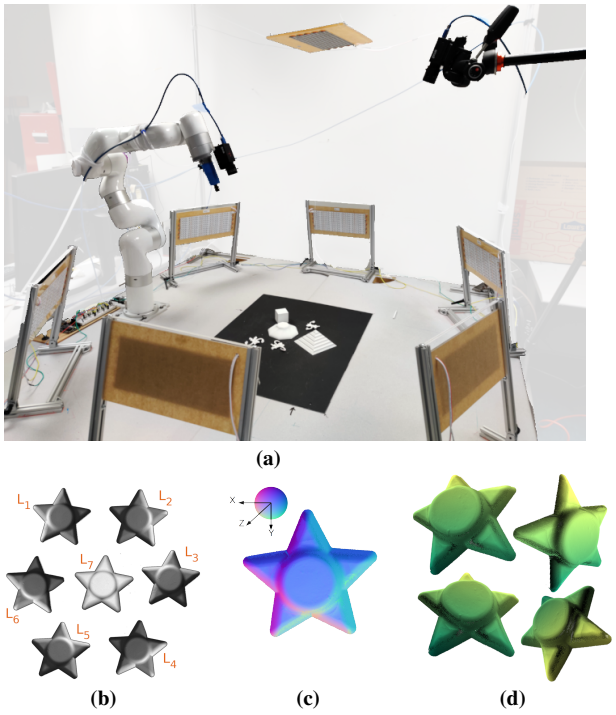


Figure 1. We demonstrate the value of controlled illumination approaches for robot manipulation workspaces. Our setup (Fig. 1a) consists of 2 machine vision cameras (C_1 mounted to the robot and C_2) overlooking a robot's workspace. The illumination of the workspace is controlled using 7 directional lights – six low angle of incidence light sources $L_{1:6}$ placed on the table and one overhead light source L_7. We capture seven images of the object (Fig. 1b), use standard photometric stereo to calculate object surface normals (color coded in Fig. 1c), and optionally derive a 3D representation of the object's surface (Fig. 1d) from calculated normals (Fig. 1c).

Our experiments focus on precise and effective sensing in a robot manipulation setup [28, 30].

To that end, this paper proposes using actively controlled illumination and tries to address a simplified version of the motivating question: How can controlled illumination be used to estimate surface properties of objects to make robotic manipulation easier? The simplifications include painting objects with matte paint to remove highlights due to specularities, and using only one color of paint (white) to make the reflectance function uniform over the object.

Active control of sensing parameters is a classical topic in robotics research – [4, 6, 7] define "active perception" as controlling environment parameters (e.g. illumination), extrinsic (e.g. sensor location) and intrinsic (e.g sensor gain, focal length) sensor parameters for better perception. Our work can be classified as a subset of active perception research where the scene illumination, camera poses, and some of the camera intrinsic parameters are actively controlled to solve the perception task at hand.

Although we will handle more general reflectance functions in future work, for this work, we assume that all objects have monochromatic Lambertian surface reflectances, i.e. they do not have specularities and changes in reflectances due to colored patches on the object. We enforce this assumption by coating all objects with matte white paint[1] when necessary. All of our objects are single rigid bodies with no articulations and when needed, we assume that we have 3D models available to us. We show that:

- Controlling illumination of a robot's workspace yields high quality information about surface normals and discontinuities – significantly better than commercially available 3D sensors for table top scale objects ($\leq$20cm^3).
- The computed normals and surface discontinuity information can aid in robot grasping tasks.
- A controlled illumination approach can help us estimate deformation of known objects, and estimate poses of known and unknown objects.

Additional details and results from our work, video demonstrations of robot tasks and summaries of this research can be found at https://arkadeepnc.github.io/projects/active_workspace/index.html.

2. Related Work

Shape from shading, introduced by Horn [23] and Woodham [49], is a classical problem in computer vision which involves obtaining the shape and surface reflectance of an object by varying either the viewing direction, or the scene illumination or both. Several versions of this problem have been proposed where the researchers have recovered the shape, reflectance and shading of the object from learned priors [9], proposed accurate methods for recovering object shapes as a collection of algebraic surfaces [50], and more recently have demonstrated highly accurate frameworks for recovering object shape, reflectance and geometry by refining multi-view RGB-D data captured by commercial sensors [33].

Although a large portion of recent work [12, 21, 34, 43] advocates learning implicit representations of objects from multiple views and then using them for robotics tasks [53, 54], there has been increasing interest in research on inferring object geometry from its interactions with controlled directional illumination. Recent work has demonstrated the capture of object surface normals and reflectances using multi-view photometric stereo [51], explicitly recovering object shape and reflectance from images taken with cameras equipped with flashes [15, 33] and have demonstrated cameras paired with an ensemble of projectors and flashes to capture scene properties [22, 40]. Other notable efforts include learning implicit scene representations from multiple directionally illuminated images [52], estimating object geometry from shadows cast by the object under directional illumination [44], and reconstructing surface depth and normals from images under directional illumination [5]. Although a large portion of these works demonstrate impressive accuracy for a selected set of objects, it is unclear how appropriate any of these are as a perception system for manipulation tasks. [27] were the first to demonstrate the use of photometric stereo as a metrically accurate sensor and smaller versions of the setup [25, 26] have been shown to be useful for several manipulation tasks. Our work draws inspiration from [27] as we demonstrate the applicability of object geometry capture for different robotic manipulation tasks using techniques from classical computer vision.

3. Methods

It is well known from the shape from shading literature [9, 23] that the measured intensity through an imaging device is a function of three major quantities – the shape and reflectance of the object and the illumination of the environment. In this work, we focus on recovering surface normals as a proxy for the object shape. We use controlled lighting in addition to ambient illumination. Further, we simplify the shape from shading problem by exclusively considering Lambertian objects – painting the objects with matte white paint so that the reflectance is known (Lambertian).

As shown in Fig. 1a, we illuminate the workspace with a low angle of incidence with approximately parallel light rays $\mathbf{l}$, emulating a light source at infinity. Objects are also indirectly illuminated by the ambient light and we model the illumination as a combination of a linear model [10]:

$$I_i^k = \frac{\rho}{\pi}\langle \mathbf{n}_k, \mathbf{l}_i \rangle \;\; \forall i \in 1..., 6 \tag{1}$$

and a quadratic model $\mathbf{M}$ using second order spherical harmonic functions [27, 36]:

$$I_i^k = {\mathbf{n}_k}^T \mathbf{M}_i \mathbf{n}_k \;\; \forall i \in 1..., 6 \tag{2}$$

[1] we use RustOleum 7790830 Flat White spray paint

At inference time, given measured pixel intensities I_i^k and albedo ρ, and per-channel linear and quadratic illumination models $\mathbf{l}_i$, $\mathbf{M}_i$, we obtain the surface normals $\mathbf{n}_k$ by inverting the models sequentially for $L_{1:6}$ in Fig. 1a, while accounting for shadows on the object. Optionally, we spatially integrate the surface normals to get a depth map of the object's surface. As our perception pipeline builds upon traditional techniques from photometric stereo, we defer a detailed discussion of our methods to the supplementary material.

4. Experiments and Results

In this section we describe our experiments on using our sensing system for common manipulation tasks. We start by quantifying the performance of our approach and then demonstrate the use of our approach in three sub-tasks related to manipulation – general purpose object picking, estimating object deformation with vision, and pose estimation. We focus the paper on the results of our experiments, deferring details of the methods to the supplementary material.

4.1. Performance of our sensing approach

4.1.1 Quality of normals versus true depth sensors

The commercial depth sensors widely used for robot manipulation tasks are fundamentally different from our sensor because the primary measurement in those sensors is the depth of the visible surface from the sensor calculated either through stereo matching (first 3 rows of Tab. 1), or through time-of-flight (rows 4, 5 of Tab. 1). For these depth sensors, we calculate the normals as spatial derivatives of the captured depth. Our sensing approach infers object surface normals from shaded images and optionally calculates a depth map that best explains the observed normals when the whole object is visible from the cameras' viewpoint. We also note that the commercial depth sensors we used (except for the D405) are designed to operate in room scale environments and are not necessarily suited for measuring surfaces of the objects we used. To focus on the quality of the estimated surface normals we measure the statistical similarity of the normals measured across all the sensors on the same object patches. We image flat and textured surfaces at two orientations – facing the camera's projection axis and at an inclination of 45° to the projection axis at the closest possible distance. We use the earth movers distance [31, 38] to compare the normals captured by our approach and the commercial sensors to the ground truth.

We also note that the stereo-based depth sensors rely on visual textures which our objects lack. To address this, we follow the recommendations for imaging textureless surfaces from [29] by projecting visible (D405) or infrared (D435, D455) patterns on the objects as applicable. Additionally, to reduce the noise in the measurements, for the stereo sensors (rows 1 through 3 in Tab. 1), we calculate the measured depth as a trimmed mean (we remove 10% of smallest and largest outliers) of all the depths at each pixel for 50 consecutive frames (acquired over 1.5 - 2 seconds).

Sensor	Flat 0°	Flat 45°	Texture 0°	Texture 45°
D455	0.12	0.12	0.27	0.12
D405	0.09	0.04	0.21	0.15
D435	0.09	0.23	0.22	0.22
L515	0.06	0.05	0.20	0.15
Kinect II	0.24	0.14	0.19	0.16
Ours	**0.02**	**0.01**	**0.18**	**0.10**

Table 1. Comparison of our approach with commercial depth sensors (lower is better). The metric value indicates the *dissimilarity of the measured normals with ground truth* which in the case of flat surfaces is related to the standard deviation of the angles of the normals with respect to the mean. Representative normal maps corresponding to the next best performing sensor have been provided in Fig. 2a along with the data captured by our sensor in those categories.

	Normals (△°)		Surface (△mm)	
Object name	$\mu(\sigma)$	$< 20°$ (%)	$\mu(\sigma)$	max
Pyramid	13.31 (12.89)	93.06	1.31 (0.93)	4.17
Star	18.85 (14.04)	86.73	0.80 (0.60)	3.65
Bent Cyl.	18.81 (15.08)	86.19	0.50 (0.36)	2.03
Octagon I	16.23 (12.53)	83.45	0.87 (0.85)	4.50
Octagon II	17.17 (12.09)	82.52	1.20 (0.90)	5.05
Spot	17.38 (10.42)	83.55	0.54 (0.48)	2.82
Bas-relief	18.24 (10.01)	76.24	0.87 (0.94)	4.26
Bunny	20.07 (14.80)	75.88	1.60 (1.27)	8.12
Text. Pyramid	19.19 (17.52)	61.75	1.74 (1.31)	7.30
Happy Budd.	29.19 (20.52)	54.29	1.54 (1.22)	7.66

Table 2. Quantitative measurements of our approach in estimating object shape. The measurements in the first set of columns are the deviations of measured normals from ground truth normals in degrees. The measurements in the second set of columns are the deviation of the reconstructed surface from the ground truth mesh in millimeters. Please visit the project website for a qualitative visualization of the results.

For the time of flight sensors (rows 4 and 5 in Tab. 1) we use a 3×3 pixel window median filter to smooth the captured depths at each pixel after temporally filtering the data using trimmed means. We present our quantitative results in Tab. 1 and our qualitative results in Fig. 2a. We outperform all the commercial sensors in imaging surfaces as normals in each category, often by significant margins. Our approach is much better at detecting that a surface is actually smooth and flat, and in measuring local surface orientation. Unsurprisingly, we also note that the D405 and L515 sensors outperform the other sensors because they were designed for (or support) close range imaging which is relevant to the task we tested the sensors on.

4.1.2 Accuracy of our approach

To quantify the accuracy of our sensor, we 3D printed a set of objects of footprint less than 12cm^2 using a standard 3D printer. The objects were imaged with our setup in Fig. 1. The surface normal quality measurement is performed by finding the object pose that aligns the measured surface normals to the ground truth surface normals generated with a renderer imaging the ground truth mesh. We used photometric stereo to calculate the normals and the method described in Sec. 4.4 to align the captured object

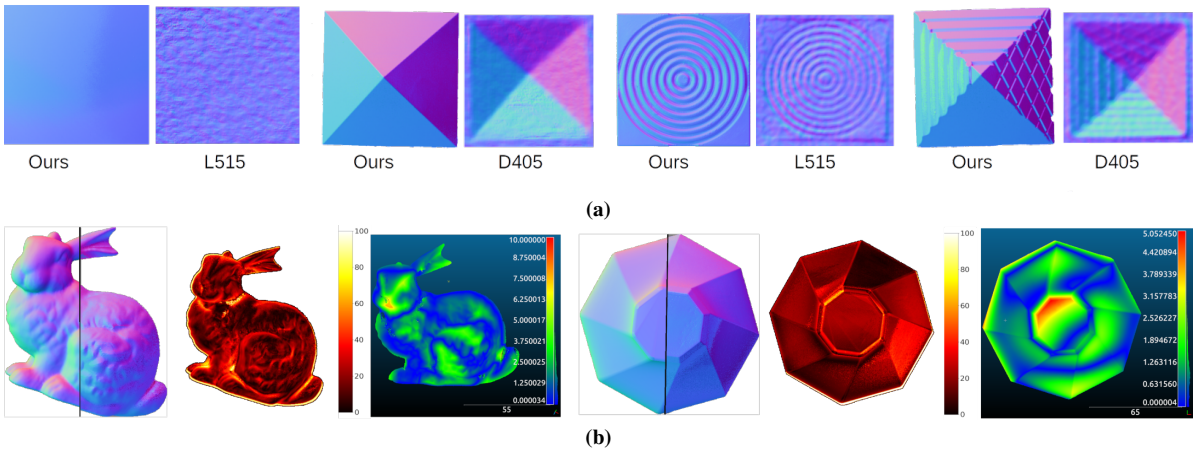


Figure 2. Our approach compared to commercial depth sensors. In Fig. 2a we compare the quality of normals computed by our sensor with the processed data from the best performing commercial depth sensor (Tab. 2). Figure 2b shows the normals of the bunny and octagon II from Tab. 2 as sets of 3 images each. In each set from left to right, we overlay the normals calculated by our approach (right of black line) on the ground truth normals (left of black line), the per pixel deviations of the normals from ground truth (in degrees) and the deviations of the measured surface from the ground truth mesh in mm.

to the ground truth mesh model. We then calculated the angle between the measured normal and the rendered ground truth normals at every pixel and reported the statistics of the angles as a quantitative measure of the quality of the normals estimated by our approach. Our pose estimation pipeline (Sec. 4.4.1) aligned the simulated and the real data up to a maximum error of 5 pixels, making the per pixel error calculation meaningful. In Tab. 2 under the label 'normal quality' we report the mean and the standard deviation of the per pixel angle error in degrees. We also report the percentage of pixels with angle error less than 20°. This metric was influenced by our observation during vacuum picking experiments (see Sec. 4.2) that with the choice of a more compliant gripper, our gripping pipeline was tolerant of surface normal errors up to 20°. Notwithstanding the errors introduced due to warps on some 3D printed models (see e.g. high error pixels around the edges of insets 2 and 5 in Fig. 2b), the quality of our normal and depth estimates are similar to classical methods [9, 50] and are marginally worse than recent deep learning based methods [5, 51, 52].

To measure the surface quality, we generated a surface depth map by spatially integrating the normals of the scene calculated using photometric stereo (more details in the supplementary materials). We then registered the integrated depth map to the ground truth mesh using point-to-plane ICP [39] and calculated the Hausdorff distance [16] between the recovered depth map and the ground truth mesh of the object. The mean, standard deviation and the maximum point-wise distance of the recovered depth map from the ground truth meshes are reported under the surface quality column of Tab. 2. For both the experiments, we observe that the objects with large planar faces or smoothly varying curvatures worked the best while, objects with undercut surfaces performed poorly – especially the happy Buddha object which had several undercut faces (see Fig. 5a). We present a qualitative result in Fig. 2b. More qualitative results can be viewed on our project website.

4.2. Controlled illumination for pickup tasks

To perform immobilizing grasps with a single suction cup vacuum gripper, we need to identify a portion of the object that is large enough for the suction cup to fit and flat enough for the suction cup to be effective. Additionally, we have to localize the point of grasp with respect to the gripper and identify the local surface orientation so that the gripper can approach along the surface normal with the face of the suction cup perpendicular to the surface to best execute the grasp. We have described pipelines to identify surface normals, demonstrated our system's performance in measuring surface discontinuities, and we have a stereo system (C_1 and C_2 in Fig. 1) for triangulating a world point in the robot's frame. In this section, we describe our pipeline to detect arbitrarily oriented flat objects that can be grasped with a suction cup gripper of a given size. We also demonstrate how low angle of incidence directional illumination can help identify and pick up thin flat objects when segmentation is difficult using color or depth contrasts with conventional sensors.

We divide the problem of picking up 3D objects into two major steps – identifying the largest geometrically flat patch in the workspace across two views and executing the robot motion to grasp the identified face. To identify the largest corresponding flat patch across the camera views, we modified the well-known CAMShift algorithm [13] to filter out image patches not graspable with a vacuum gripper.

With geometrically corresponding grasping locations identified across two views, we use camera poses and intrinsics of the two views C_1 and C_2 to triangulate the position of the gripper in the robot's coordinate frame. For the orientation of the gripper – we calculate the normal at the grasp location by averaging the normals at the identified grasp location in the two views. We then generate and exe-

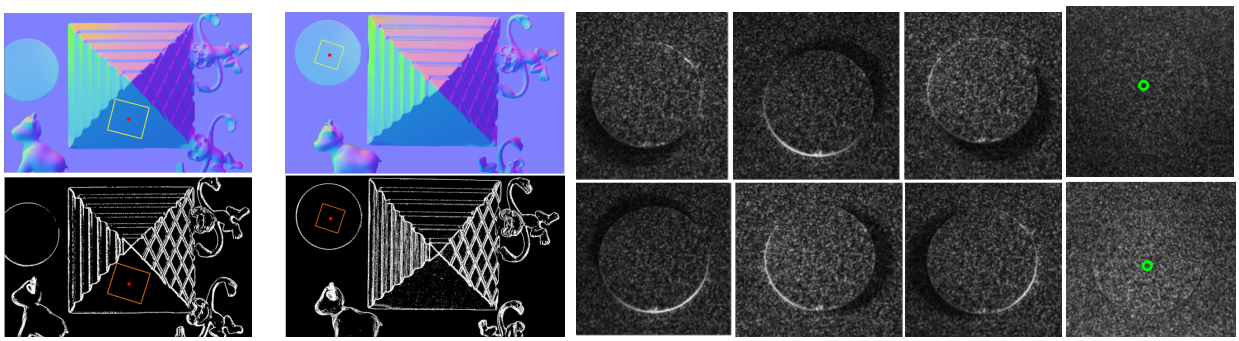


Figure 3. Our pipeline for grasping objects in a robot's workspace. Figures 3a and 3b respectively show two distinct grasps selected by the camera C_1 along X and Y directions respectively. On the top rows of Figs. 3a and 3b we show the detected grasps projected on the scene normals and, on the bottom row, we show the detected grasps projected onto the occlusion edge pixels of the scene. We note that our selected grasps do not have occlusion edges. Figures 3c and 3d shows our pipeline for picking up thin objects in absence of obvious depth or texture segmentation cues. Figure 3c shows the images of a plastic disc (50mm diam., 3.15mm thick) captured with C_1 placed vertically above the disc, using the directional lights $L_{1:6}$. Figure 3d shows our detection of the center of the discs overlaid on the image captured with cameras C_1 (top) and C_2 (bottom) with L_7 and ambient light illuminating the scene.

cute a minimum jerk trajectory [19] which moves the robot to make contact with the selected object, picks the object up and, places it in a bin in the robot's workspace.

For picking up thin objects without color contrast following Raskar et al. [37], we observe that image brightness variations due to occlusion are more prominent than brightness differences due to color texture. We use this intuition to identify thin and flat objects from the background. For this experiment we pick a plastic disc of 50mm diameter, and 3.15mm thickness with a random pattern (pixel values sampled uniformly between 0 to 255) printed on the surface of the disc from a flat plane with the same visual texture. As is obvious from Figs. 3c and 3d, there isn't enough intensity or color texture to segment the object from the background. However, looking along the light direction (see Fig. 3c), we note that the shadows due to occlusion are more prominent than the visual textures on the object and the background – as imaged in Figs. 3c and 3d.

This shadow cue was sufficient for our edge detection procedure (see supplementary materials for details) to identify occlusion edge pixels in the scene. We then estimated the center of the disc by fitting a minimum enclosing circle [35] to the occlusion edge pixels across the two views C_1 and C_2. We visualize this in Fig. 3d – the centers of the circles have been projected onto the images captured by the two views with the ambient and overhead light (L_7 in Fig. 1a). To pick up the disc, we approached the center of the disc along the surface normal of the tabletop by triangulating the object from the two views in Fig. 3d. None of the commercial depth sensors we used (see Tab. 1) detected the disc reliably.

We performed 35 grasp experiments with 3D objects and ten experiments with flat textured objects (5 with the disc, and 5 with an irregular octagon inscribed by a 50mm diameter circle and of 3.15mm thickness). Movies of our experiments can be viewed on the project website. Across all the experiments we failed four times while picking up the textured pyramid (Fig. 3a) due to the gripper failing to attach or colliding with the object due to triangulation errors.

4.3. Measuring deformation

Our approach can be used to perceive the deformation of objects by tracking the change of local normals on the object's surface. In this section, we describe our method for measuring the buckling deformation of a 0.76 mm thick rectangular PVC card of size 86×54mm. We measure the deformation in a typical analysis-by-synthesis fashion – we explain the change in the surface normals of the deformed card by calculating the deformation of the card geometry.

Figure 4b shows a qualitative result of our pipeline – the first image shows the reconstructed shape before the onset of buckling and the second image shows the reconstructed mesh overlaid on the deformed object. In both cases, we show the local change in surface normals of the reconstructed shape as the color of the mesh. An external view of the objects is also provided as the inset to the images. During our experiments we observed that skewed viewing directions of the cameras C_1 or C_2, e.g. the inset views in Fig. 4b, significantly reduced the performance of our pipeline due to the decrease in the effective number of pixels imaging the object across the two views. This led us to generally use C_1 and C_2 to capture frontal views. More qualitative results of our experiments can be viewed on the project website. Detailed explanation of our pipeline can be found in the supplementary materials.

To verify that we indeed captured the physical process behind the card buckling under vertical loads, we compared our predictions with theoretical values calculated using solid mechanics. The theoretical data showed a linear trend in the mean change in surface normals and the vertical deformation of the card. A line with a slope of 1.78° and offset of 0.114 fits the theoretical data with a r^2 score of 0.975. We repeated the experiment of deforming the card five times with five different PVC cards and different camera positions (C_1 and C_2 in Fig. 4a) and obtained 23 data points as shown in Fig. 4c with five different markers cor-

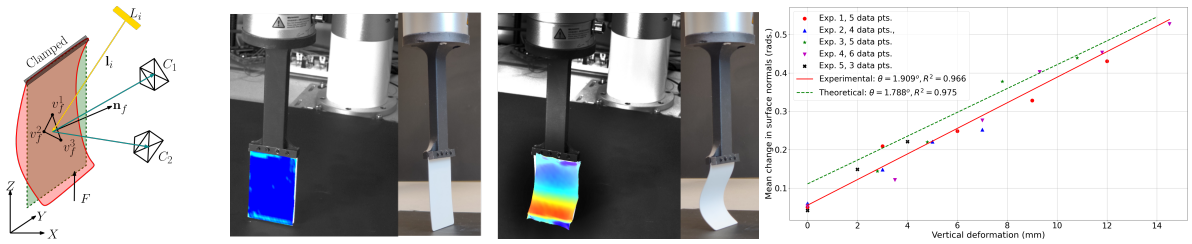


Figure 4. Our results of measuring the deformation of known objects. Figure 4a shows a schematic diagram of our procedure, Fig. 4b shows the estimated shape of the card overlaid on a camera image. The insets of the images in Fig. 4b provide an external view of the state of the cards. Figure 4c summarizes the quantitative results of result of our experiment to show we were able to estimate the physical process behind the deformation of the card.

Method	Scale	Error $\mu(\sigma)$	Max	Sample size
VIRDO [46]	89×56	1.12 (- - -)	–	5.6k
VIRDO++ [47]	89×56	1.04 (0.35)	–	5.6k
Ours	86×54	0.78 (0.52)	4.57	100k

Table 3. Comparison of our method of measuring deformation (chamfer distance) against baselines [46, 47]. Our experiment (single view) was carried out in simulation because we do not have access to a sensor accurate enough to measure ground truth deformations, and we are comparing the performance of Wi and colleagues' results during inference on simulated data with a single camera view. Our experiment captures the deformed card at a simulated endpoint deflection of 12 mm corresponding to a mean change in surface normals of 0.41 radians. All length measurements are in mm. and all the experiments assume full visibility of the object.

responding to the experiments. The experimental data also showed a linear trend – a line with slope of 1.9°, offset of 0.055 fit the data with a r^2 score of 0.966. Disregarding the difference in the offset due to the baseline noise in our measurements and the slight violation of a geometric boundary condition due to slipping of the card at the base, from Fig. 4c we can conclude that our pipeline is repeatable and can capture the physical process of the buckling deformation of the card due to vertical loads.

We note that in this work we exclusively use vision. We do not factor in the force that is causing the deformation in the object. In related work by Wi et al. [46, 47] the authors combine visual measurements with forces measured by a force torque sensor coupled to the deforming object to infer its deformation. For objects of similar scale, we compare our performance with the works of Wi and colleagues in Tab. 3. Although we perform better in reconstructing shapes, we require nearly full visibility of the object in both views which is not a limitation for Wi et al. in [47].

We also note that our "analysis-by-synthesis" procedure works better than integrating the captured normal map (see supplementary for details) to reconstruct the surface of the card. This is due to the strong view dependence of the shape reconstruction, which prevents us from trivially incorporating multiple views to reconstruct the bent object. Using two camera views and an initial mesh of the object to predict the shape of the deforming card makes our pipeline more robust to shadows and slight occlusions in camera views.

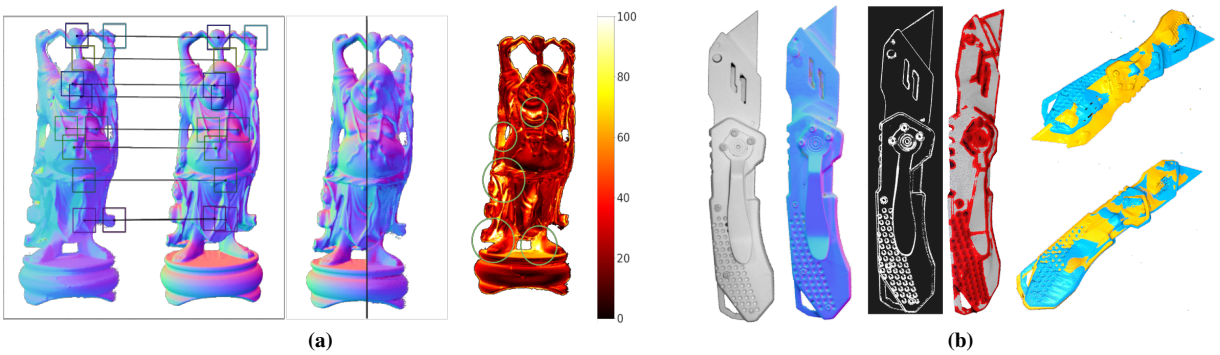
4.4. Controlled illumination for localization

4.4.1 Estimating poses of known objects

When an object model is available, we estimate its pose by aligning the measured surface normals, depth edges, and object silhouettes of the object with their simulated counterparts. We improve on commonly employed pipelines in the pose estimation literature [14, 24, 32] by generating good initial pose estimates through discovery of object patch correspondences between the observed and simulated surface normals (inset 1 of Fig. 5a). We describe our method in detail in the supplementary material.

We present a qualitative result in Fig. 5a – the second inset overlays the normal images with measured normals $\mathbf{N}_R$ on the left of the black line and rendered normals $\mathbf{N}_S$ on the right. In the third inset of Fig. 5a, we visualize our per-pixel pose estimation costs overlaid on the observed image silhouette $\mathbf{M}_S$ on an arbitrary scale between 1 to 100. We note that the under-cut parts of the geometry, highlighted by green circles, have large local costs, but we do not incur large costs due to silhouette misalignment or scaling as seen by the absence of high cost patches at the object edges or background. For all the objects tested, our multi-scale and multi-modal pose estimation pipeline reliably converges to the correct pose within a 5.5 pixel error (7 px/mm) even with high local errors in the measured data – the model in Fig. 5a has ∼50% of pixels with erroneous normals ($> 20°$ deviation from ground truth).

We present quantitative measurements of our pose estimation approach for all the objects in Tab. 2 and Tab. 4. Our experiments were done with C_1 imaging a 227mm × 129mm area at a resolution of 7 pixels/mm. We placed each of the objects at a random position and orientation and used the pipeline described in this section to align the 3D model to the observed data. After the alignment, we overlaid the simulated and measured data and manually measured the pixel misalignment between the real and simulated data around the object edges. We repeated this experiment five times for each object and report the mean and standard deviations of the pixel misalignment in Tab. 4.



(a) (b)

Figure 5. Estimating object pose in the robot workspace. Figure 5a shows our steps in estimating object poses when a 3D model of the object is available apriori. From left to right in Fig. 5a we calculate the pixel-wise correspondences between captured data and the available geometry, overlay the captured surface normals (right of black line) onto the ground truth object normals (left of black line) and show our pixel-wise pose estimation costs about an arbitrary scale. The green circles indicate areas with high local costs due to errors in estimating normals. Figure 5b shows our pipeline for tracking object pose when a 3D model is not available a priori. From left to right, we show our captured data, calculated normals, calculated depth edges, calculated 3D representation of the surface with salient point-features overlaid on the points in red and, initial and final stages of our estimation of the object's pose.

Object	Bounding Box (mm)	Px. Errors $\mu(\sigma)$
Pyramid	$100 \times 100 \times 55$	2.50 (1.70)
Star	$110 \times 110 \times 15$	3.62 (1.03)
Bent Cylinder	$116 \times 48 \times 13$	5.50 (2.12)
Octagon I	$100 \times 100 \times 29$	3.35 (0.88)
Octagon II	$100 \times 100 \times 29$	2.85 (1.83)
Spot	$52 \times 15 \times 53$	1.50 (0.23)
Bas-relief	$83 \times 110 \times 7$	0.80 (0.20)
Bunny	$93 \times 94 \times 43$	4.06 (1.19)
Text Pyramid	$100 \times 100 \times 55$	4.02 (0.96)
Happy Buddha	$115 \times 45 \times 37$	3.10 (0.69)

Table 4. Performance of pose estimation for known objects. All the experiments were done with a resolution of ∼7 pixels/mm. Qualitative results for the objects happy Buddha, bunny and octagon II can be found in Fig. 5a and the first and fourth insets of Fig. 2b respectively. More qualitative results can be found on the project website.

4.4.2 Estimating the pose of unknown objects

If a 3D model of an object is not available, the pose estimation problem, defined classically, is ill-posed. In those cases, we can estimate the object's change in pose through rigid registration of the 3D representations of the same object as it moves. For our 3D representation, we choose the point cloud obtained by integrating the surface normal maps, factoring in the camera intrinsics (see supplementary for details). As noted before, this representation is not metrically correct for parts of the object that are not fully visible by the camera (e.g. full profile of the belt clip of the knife in Fig. 5b), however, the relative surface depth changes and the overall scale of the object are captured accurately which lets us estimate the change in pose between two measurements of the same object. We pictorially describe our pipeline for tracking unknown objects in Fig. 5b, which involves capturing the image of the object, measuring its surface normals and depth edges, generating a point cloud of the observed surface and registering two instances of the point clouds to obtain the change in pose. We describe our pipeline in detail in the supplementary material.

Object	Bounding Box (mm)	$\triangle X$(mm)	$\triangle Y$(mm)	$\triangle\theta(^\circ)$
Knife	$172 \times 30 \times 20$	1.02	1.66	0.36
Monkey	$46 \times 64 \times 8$	1.67	0.97	0.22
Circuit	$13 \times 21 \times 33$	3.28	2.40	1.28
IO shield	$120 \times 39 \times 21$	1.48	1.86	0.25

Table 5. Performance of our tracking pipeline for unknown objects. All the experiments were done with a resolution of approximately 7 pixels/mm. Qualitative results for the knife is shown in Fig. 5b, to view results of the other objects, please visit the project website.

To evaluate our pipeline for estimating pose changes of unknown objects, we imaged four objects of different scales twice, while introducing a known pose perturbation between two measurements and recovered the pose perturbation using the method described in this section. We repeated this experiment six times for each of the objects in Tab. 5 and report our pipeline's uncertainty in recovering the pose perturbations. We present our quantitative results in Tab. 5 – our approach has a tracking uncertainty of about 2° in planar rotation and about 4 mm in translation.

Counter-intuitively, we also noted that generating a mesh of the object from the captured 3D representation and then using the mesh in the pipeline discussed in Sec. 4.4.1 actually led to poorer pose estimation because the successive processing steps (normal calculation, integration, and meshing) reduced the quality of the mesh input. Since the 3D representation generated is strongly dependent on the viewpoint of the camera during the experiment meant that the 3D model was metrically incorrect for novel views.

4.5. Application to generic objects

Our previous experiments focused on Lambertian objects with diffuse, white paint in calibrated environments. Extending these techniques to more general objects is possible using off-the-shelf methods for intrinsic image decomposition. Intrinsic image decomposition separates natural color images into a reflectance image and a shading image.

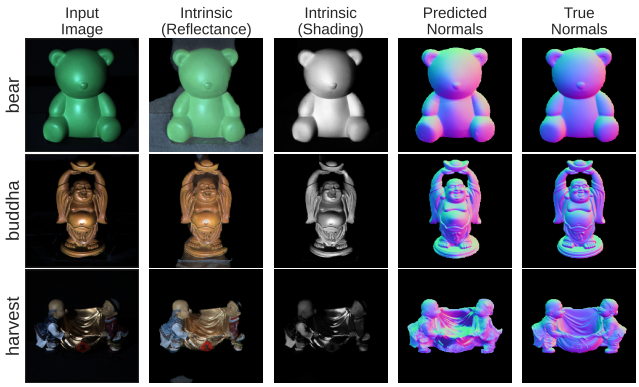


Figure 6. Our photometric stereo on colored, non-Lambertian objects using intrinsic image decomposition [17] on a dataset [42]. Sec. 4.5.

These shading images can be used as input to any typical photometric stereo method, as they estimate the isolated impact of lighting across the surface. There are many approaches to intrinsic image decomposition, including classic priors [8] and data-driven learning methods [11, 20]. In our experiments, we use a recent method called PIE-Net [17] to perform the intrinsic image decomposition.

We show the results of this approach on three examples from the DiLiGenT [42] dataset, which contains several objects that include not only color but also general non-Lambertian materials. We show the results of a simple pipeline, using the pretrained PIE-Net model to obtain shading images. Photometric stereo is solved using the global, linear model that assumes distant point light sources and Lambertian shading. For each scene, DiLiGenT provides a single view of the scene with 95 different illumination conditions. We used 30-40 lighting directions for our reconstructions, each at PIE-Net's 256^2 image resolution.

Our results are shown in Fig. 6. While this pipeline produced good results for some objects (bear, buddha), it struggled with the glossy, dark, and self-occluding harvest object. Nonetheless, considering that our pipeline used an off-the-shelf network and classic photometric stereo (without explicit handling of shadows or specularities), the results are promising. While this approach is far from the metric quality of our robotics setup in Secs. 4.1.2, 4.4.1 and 4.4.2, we believe this quality of normals is potentially promising for vaccum gripper tasks such as those in Sec. 4.2 or the deformation estimates in Sec. 4.3. While we focused on the high end of precision in photometric stereo, this type of off-the-shelf pipeline enables generalization to generic objects, and produces potential utility in less demanding tasks.

5. Discussion and Future Work

Having a Lambertian reflectance requirement for our objects is restrictive and can be a barrier for our methods to apply to many manipulation tasks – future work can build on Sec. 4.5 to enroll general objects into our pipeline. During this work, we observed that although we do well in inferring the cumulative shape of the object we often have high local errors – see e.g. the second inset of Fig. 2b, and third inset of Fig. 5a. We believe that these local errors are due to shadows that could not be resolved by our proposed method of inferring shape at a single pixel level with lights that are not co-incident with the camera. Further research is required on multiplexing the illumination sources to reduce the effect of shadows and to determine a better combination of light and camera locations. Lights collocated with cameras and on the robot workspace will possibly address some of the limitations of our work. Further research is also required for selecting alternative object representations. Current literature indicates that a triangle-based representation [33] or locally smooth patch-based representations [50] may work but will need a plethora of hand-tuned regularizers, hyperparameters and a significant amount of computational effort to converge to a meaningful representation. Volumetric representations [21, 34] have certain advantages over patch-based representations in the context of robotic manipulation. Integrating volumetric representations with a robot workspace scaled controlled illumination approach is future work. In this work, we achieve a higher fidelity of measurement than some commercial depth sensors by imposing priors on object reflectances and controlling illumination. A natural extension of the current work would be to augment the performance of a commercial depth sensor by using it in conjunction with our approach.

6. Conclusions

In this work we demonstrated the application of classical techniques from photometric stereo to robotic manipulation through a robot workspace scaled controlled illumination system. We showed that, by enforcing a reflectance prior on the objects, reasoning about observed object intensities conditioned on the direction of illumination can yield accurate surface normals and identify surface depth discontinuities with very little computation. We also showed that the normals captured by our approach are significantly better than the ones derived from measurements with commercial depth sensors and we also evaluated the accuracy of our approach in capturing surface depth and normals. With the surface representations generated using our approach, we demonstrated three common manipulation tasks – picking up objects of arbitrary shape with a single point vacuum gripper, estimating bending deformation of a known object and estimating poses of Lambertian objects.

References

[1] Ensenso XR series scanners. 1
[2] Photoneo PhoXi3D scanners. 1
[3] John Aloimonos. Shape from texture. *Biological Cybernetics*, 58(5):345–360, 1988. 1
[4] John Aloimonos, Isaac Weiss, and Amit Bandyopadhyay. Active vision. *International Journal of Computer Vision*, 1(4):333–356, 1988. 2
[5] Yacov Hel-Or Asaf Karnieli, Ohad Fried. DeepShadow: Neural shape from shadows. In *ECCV*, 2022. 2, 4
[6] Ruzena Bajcsy. Active perception. *Proceedings of the IEEE*, 76(8):966–1005, 1988. 2

[7] Ruzena Bajcsy, Yiannis Aloimonos, and John K Tsotsos. Revisiting active perception. *Autonomous Robots*, 42(2):177–196, 2018. 2

[8] Jonathan T. Barron and Jitendra Malik. Intrinsic scene properties from a single rgb-d image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013. 8

[9] Jonathan T Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(8):1670–1687, 2014. 2, 4

[10] Svetlana Barsky and Maria Petrou. The 4-source photometric stereo technique for three-dimensional surfaces in the presence of highlights and shadows. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(10):1239–1252, 2003. 2

[11] Anil S. Baslamisli, Hoang-An Le, and Theo Gevers. CNN Based Learning Using Reflection and Retinex Models for Intrinsic Image Decomposition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 8

[12] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T Barron, Ce Liu, and Hendrik Lensch. NeRD: Neural reflectance decomposition from image collections. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12684–12694, 2021. 2

[13] Gary R Bradski. Computer vision face tracking for use in a perceptual user interface. 1998. 4

[14] Arkadeep Narayan Chaudhury, Timothy Man, Wenzhen Yuan, and Christopher G Atkeson. Using Collocated Vision and Tactile Sensors for Visual Servoing and Localization. *IEEE Robotics and Automation Letters*, 7(2):3427–3434, 2022. 1, 6

[15] Ziang Cheng, Junxuan Li, and Hongdong Li. WildLight: In-the-wild Inverse Rendering with a Flashlight. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4305–4314, 2023. 2

[16] Paolo Cignoni, Claudio Rocchini, and Roberto Scopigno. Metro: measuring error on simplified surfaces. In *Computer Graphics Forum*, volume 17, pages 167–174. Wiley Online Library, 1998. 4

[17] Partha Das, Sezer Karaoglu, and Theo Gevers. PIE-Net: Photometric Invariant Edge Guided Network for Intrinsic Image Decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19790–19799, 2022. 8

[18] Elliott Donlon, Siyuan Dong, Melody Liu, Jianhua Li, Edward Adelson, and Alberto Rodriguez. Gelslim: A high-resolution, compact, robust, and calibrated tactile-sensing finger. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1927–1934. IEEE, 2018. 1

[19] Tamar Flash and Neville Hogan. The coordination of arm movements: an experimentally confirmed mathematical model. *Journal of Neuroscience*, 5(7):1688–1703, 1985. 5

[20] David Forsyth and Jason J Rock. Intrinsic image decomposition using paradigms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7624–7637, 2021. 8

[21] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance Fields Without Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5501–5510, 2022. 2, 8

[22] Tomoaki Higo, Yasuyuki Matsushita, Neel Joshi, and Katsushi Ikeuchi. A hand-held photometric stereo camera for 3-d modeling. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1234–1241. IEEE, 2009. 2

[23] Berthold KP Horn. Shape from shading: A method for obtaining the shape of a smooth opaque object from one view. 1970. 2

[24] Marco Imperoli and Alberto Pretto. D^2CO: Fast and Robust Registration of 3D Textureless Objects Using the Directional Chamfer Distance. In *International Conference on Computer Vision Systems*. Springer, 2015. 6

[25] GelSight Inc. GelSight Mini. [Online; accessed 09-Dec-2022]. 2

[26] GelSight Inc. GelSight Mobile. [Online; accessed 09-Dec-2022]. 2

[27] Micah K Johnson, Forrester Cole, Alvin Raj, and Edward H Adelson. Microgeometry capture using an elastomeric sensor. *ACM Transactions on Graphics (TOG)*, 30(4):1–8, 2011. 2

[28] Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, et al. Scalable deep reinforcement learning for vision-based robotic manipulation. In *Conference on Robot Learning*, pages 651–673. PMLR, 2018. 2

[29] Leonid Keselman, John Iselin Woodfill, Anders Grunnet-Jepsen, and Achintya Bhowmik. Intel realsense stereoscopic depth cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–10, 2017. 3

[30] Oliver Kroemer, Scott Niekum, and George Konidaris. A review of robot learning for manipulation: Challenges, representations, and algorithms. *The Journal of Machine Learning Research*, 22(1):1395–1476, 2021. 2

[31] Haibin Ling and Kazunori Okada. An efficient earth mover's distance algorithm for robust histogram comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(5):840–853, 2007. 3

[32] Ming-Yu Liu, Oncel Tuzel, Ashok Veeraraghavan, Yuichi Taguchi, Tim K Marks, and Rama Chellappa. Fast object localization and pose estimation in heavy clutter for robotic bin picking. *The International Journal of Robotics Research*, 31(8):951–973, 2012. 6

[33] Fujun Luan, Shuang Zhao, Kavita Bala, and Zhao Dong. Unified shape and svbrdf recovery using differentiable Monte Carlo rendering. In *Computer Graphics Forum*, volume 40, pages 101–113. Wiley Online Library, 2021. 2, 8

[34] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2, 8

[35] OpenCV. minEnclosingCircle(): Open Source Computer Vision Library, 2022. 5

[36] Ravi Ramamoorthi and Pat Hanrahan. An efficient representation for irradiance environment maps. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, pages 497–500, 2001. 2

[37] Ramesh Raskar, Kar-Han Tan, Rogerio Feris, Jingyi Yu, and Matthew Turk. Non-photorealistic camera: depth edge detection and stylized rendering using multi-flash imaging. *ACM Transactions on Graphics (TOG)*, 23(3):679–688, 2004. 5

[38] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000. 3

[39] Szymon Rusinkiewicz and Marc Levoy. Efficient variants of the ICP algorithm. In *Proceedings of the Third International Conference on 3-D Digital Imaging and Modeling*, pages 145–152. IEEE, 2001. 4

[40] Carolin Schmitt, Simon Donne, Gernot Riegler, Vladlen Koltun, and Andreas Geiger. On joint estimation of pose, geometry and svbrdf from a handheld scanner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3493–3503, 2020. 2

[41] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise View Selection for Unstructured Multi-View Stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 1

[42] Boxin Shi, Zhe Wu, Zhipeng Mo, Dinglong Duan, Sai-Kit Yeung, and Ping Tan. A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3707–3716, 2016. 8

[43] Pratul P Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T Barron. NeRV: Neural reflectance and visibility fields for relighting and view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7495–7504, 2021. 2

[44] Kushagra Tiwary, Tzofi Klinghoffer, and Ramesh Raskar. Towards learning neural representations from shadows. *arXiv preprint arXiv:2203.15946*, 2022. 2

[45] Dor Verbin and Todd Zickler. Toward a Universal Model for Shape From Texture. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[46] Youngsun Wi, Pete Florence, Andy Zeng, and Nima Fazeli. VIRDO: Visio-tactile implicit representations of deformable objects. *arXiv preprint arXiv:2202.00868*, 2022. 6

[47] Youngsun Wi, Andy Zeng, Pete Florence, and Nima Fazeli. VIRDO++: Real-World, Visuo-tactile Dynamics and Perception of Deformable Objects. *arXiv preprint arXiv:2210.03701*, 2022. 6

[48] Wikipedia. Stucco, 2022. [Online; accessed 09-July-2022]. 1

[49] Robert J Woodham. Photometric method for determining surface orientation from multiple images. *Optical Engineering*, 19(1):139–144, 1980. 2

[50] Ying Xiong, Ayan Chakrabarti, Ronen Basri, Steven J Gortler, David W Jacobs, and Todd Zickler. From shading to local shape. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(1):67–79, 2014. 2, 4, 8

[51] Wenqi Yang, Guanying Chen, Chaofeng Chen, Zhenfang Chen, and Kwan-Yee K Wong. PS-NeRF: Neural inverse rendering for multi-view photometric stereo. In *European Conference on Computer Vision*, pages 266–284. Springer, 2022. 2, 4

[52] Wenqi Yang, Guanying Chen, Chaofeng Chen, Zhenfang Chen, and Kwan-Yee K Wong. S3-NeRF: Neural Reflectance Field from Shading and Shadow under a Single Viewpoint. *arXiv preprint arXiv:2210.08936*, 2022. 2, 4

[53] Lin Yen-Chen, Pete Florence, Jonathan T Barron, Tsung-Yi Lin, Alberto Rodriguez, and Phillip Isola. NeRF-Supervision: Learning Dense Object Descriptors from Neural Radiance Fields. *arXiv preprint arXiv:2203.01913*, 2022. 2

[54] Lin Yen-Chen, Pete Florence, Jonathan T Barron, Alberto Rodriguez, Phillip Isola, and Tsung-Yi Lin. iNeRF: Inverting neural radiance fields for pose estimation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1323–1330. IEEE, 2021. 2

[55] Wenzhen Yuan, Siyuan Dong, and Edward H Adelson. Gelsight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors*, 17(12):2762, 2017. 1