Depth from Shading, Defocus, and Correspondence Using Light-Field Angular Coherence

Michael W. Tao
mtao@berkeley.edu

Pratul P. Srinivasan
pratul@berkeley.edu

Jitendra Malik
malik@eecs.berkeley.edu

University of California, Berkeley

Szymon Rusinkiewicz
smr@cs.princeton.edu
Princeton University

Ravi Ramamoorthi
ravir@cs.ucsd.edu
University of California, San Diego

Abstract

Light-field cameras are now used in consumer and industrial applications. Recent papers and products have demonstrated practical depth recovery algorithms from a passive single-shot capture. However, current light-field capture devices have narrow baselines and constrained spatial resolution; therefore, the accuracy of depth recovery is limited, requiring heavy regularization and producing planar depths that do not resemble the actual geometry. Using shading information is essential to improve the shape estimation. We develop an improved technique for local shape estimation from defocus and correspondence cues, and show how shading can be used to further refine the depth.

Light-field cameras are able to capture both spatial and angular data, suitable for refocusing. By locally refocusing each spatial pixel to its respective estimated depth, we produce an all-in-focus image where all viewpoints converge onto a point in the scene. Therefore, the angular pixels have ***angular coherence****, which exhibits three properties:* ***photo consistency****,* ***depth consistency****, and* ***shading consistency****. We propose a new framework that uses angular coherence to optimize depth and shading. The optimization framework estimates both general lighting in natural scenes and shading to improve depth regularization. Our method outperforms current state-of-the-art light-field depth estimation algorithms in multiple scenarios, including real images.*

1. Introduction

Light-fields [15, 25] can be used to refocus images [27]. Light-field cameras also enable passive and general depth estimation [32, 33, 35]. A key advantage is that multiple cues, such as defocus and correspondence can be obtained from a single shot [32]. Our main contribution is integrating a third cue: shading, as shown in Fig. 1.

We make the common assumption of Lambertian surfaces under general (distant) direct lighting. We differ from

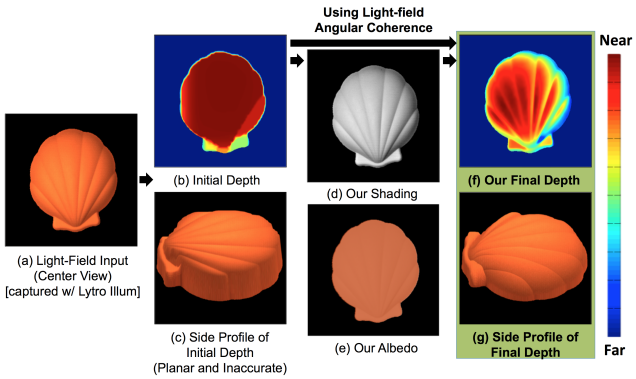


Figure 1. Light-field Depth Estimation Using Shading, Defocus, and Correspondence Cues. *In this work, we present a novel algorithm that estimates shading to improve depth recovery using light-field angular coherence. Here we have an input of a real scene with a shell surface and a camera tilted slightly toward the right of the image (a). We obtain an improved defocus and correspondence depth estimation (b,c). However, because local depth estimation is only accurate at edges or textured regions, depth estimation of the shell appears regularized and planar. We use the depth estimation to estimate shading, which is S (d), the component in $I = AS$, where I is the observed image and A is the albedo (e) With the depth and shading estimations, we can refine our depth to better represent the surface of the shell (f,g). Throughout this paper, we use the scale on the right to represent depth.*

shape from shading from single images, by exploiting the full angular data captured by the light-field. Our algorithm is able to use images captured with the Lytro and Lytro Illum cameras. We compare our results against the Lytro Illum software and other state of the art methods (Fig. 7), demonstrating that our results give accurate representations of the shapes captured. Upon publication, we will release our source code and dataset.

Shape from shading is a heavily under-constrained problem and usually only produces accurate results when fairly accurate initial depths are available for subsequence shading-based optimization [3, 6, 9, 21, 39]. Unfortunately, captured light-field data typically does not provide such information, because of the narrow baseline and limited reso-

lution. The depth estimation performs poorly, especially in smooth surfaces where even sparse depth estimation is either inaccurate or non-existent. Moreover, because we use a consumer camera to capture real world images, noise poses a large problem in estimating shading.

We represent the 4D light-field data as an epipolar image (EPI) with spatial pixel coordinates (x, y) and their angular pixel coordinates (u, v). When refocused to the correct depth, the angular pixels corresponding to a single spatial pixel represent viewpoints that converge on one point on the scene, exhibiting *angular coherence*. Angular coherence means the captured data would have **photo consistency**, **depth consistency**, and **shading consistency**, shown in Fig. 2. We extend this observation from Seitz and Dyer [30] by finding the relationship between refocusing and achieving angular coherence (Fig. 2). The extracted central pinhole image from the light-field data helps us enforce the three properties of angular coherence. We then exploit these three properties to improve the depth from defocus and correspondence introduced by Tao et al. [32]. The angular coherence and accurate confidence measures provide robust constraints to estimate shading, previously not possible with low-density depth estimation.

In this paper, our main contributions are

1. Analysis of refocusing and angular coherence (Sec. 3). We show the relationship between refocusing a light-field image and angular coherence to formulate new depth measurements and shading estimation constraints.

2. Depth estimation and confidence metric (Sec. 4.1). We formulate a new local depth algorithm to perform correspondence and defocus using angular coherence.

3. Shading estimation constraints (Sec. 5.1 and 5.2). We formulate a new shading constraint, that uses angular coherence and a confidence map to exploit light-field data.

4. Depth refinement with the three cues (Sec. 5.3). We design a novel framework that uses shading, defocus, and correspondence cues to refine shape estimation.

2. Previous Work

2.1. Shape from Shading and Photometric Stereo

Shape from shading has been well studied with multiple techniques. Extracting geometry from a single capture [18, 40] was shown to be heavily under constrained. Many works assumed known light source environments to reduce the under constrained problem [11, 12, 17, 40]; some use partial differential equations, which require near ideal cases with ideal capture, geometry, and lighting [8, 24]. In general, these approaches are especially prone to noise and require very controlled settings. Recently, Johnson and Adelson [22] described a framework to estimate shape under natural illumination. However, the work requires a known reflectance map, which is hard to obtain. In our work, we focus on both general scenes and unknown lighting, without requiring geometry or lighting priors. To relax lighting constraints, assumptions about the geometry can be made

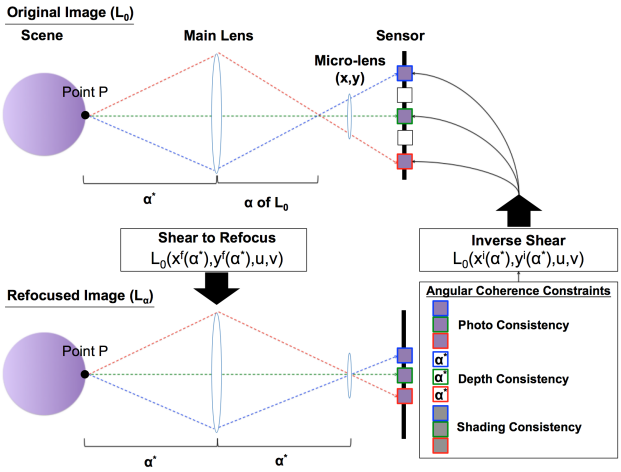


Figure 2. Angular Coherence and Refocusing. *In a scene where the main lens is focused to point P with a distance α^* from the camera, the micro-lenses enable the sensor to capture different viewpoints represented as angular pixels as shown on the bottom. As noted by Seitz and Dyer [30], the angular pixels exhibit angular coherence, which gives us photo, depth, and shading consistency. In our paper, we extend this analysis by finding a relationship between angular coherence and refocusing, as described in Sec. 3. In captured data, pixels are not guaranteed to focus at α (shown on the top). Therefore, we cannot enforce angular coherence on the initial captured light-field image. We need to shear the initial light-field image using Eq. 1 from Sec. 3, use the angular coherence constraints from Sec. 3, and remap the constraints back to the original coordinates using Eq. 7 from Sec. 3.1.*

such as faces [7, 31] or other data-driven techniques [3]. The method by Barron and Malik [1, 2] works for real-world scenes and recovers shape, illumination, reflectance, and shading from an image. However, many constraints are needed for both geometry and illumination. In our framework, we do not need any priors and have fewer constraints.

A second set of works focuses on using photometric stereo [4, 10, 12, 17, 36, 37]. These works are not passive and require the use of multiple lights and captures. In contrast, shape from shading and our technique just require a single capture.

2.2. Shape from Depth Cameras and Sensors

More recent work has been done using Kinect data [13]. Barron and Malik [3] introduce SIRFS that reconstructs depth, shading, and normals. However, the approach requires multiple shape and illumination priors. Moreover, the user is required to assume the number of light sources and objects in the scene. Chen and Koltun [9] introduce a more general approach to perform intrinsic image decomposition. However, the method does not optimize depth and, given sparse input depth with poor normal estimations at smooth surfaces, their shading estimation is poor and unsuitable for refining depth. Other works [26, 38] introduce an efficient method to optimize depth using shading information. The limitations of these approaches are that they require very dense and accurate depth estimation, achieved by

active depth cameras. Even in non-textured surfaces, these active systems provide meaningful depth estimations. With passive light-field depth estimation, the local depth output has no or low-confidence data in these regions.

2.3. Shape from Light-Fields and Multi-View Stereo

Since light-fields and multi-view stereo are passive systems, these algorithms struggle with the accuracy of depth in low-textured regions [23, 29, 32, 33, 35] because they rely on local contrast, requiring texture and edges. With traditional regularizers [20] and light-field regularizers, such as one proposed by Wanner et al. [14], depth labeling is planar in these low-textured regions. In this paper, we show how the angular coherence of light-field data can produce better 1) depth estimation and confidence levels, and 2) regularization. Van Doorn et al. [34] explain how light-fields provide useful shading information and Hasinoff and Kutulakos [16] explain how focus and aperture provide shape cues. We build on Tao et al. [32] and these observations to improve depth estimation from defocus and correspondence cues, and additionally incorporate shading information.

3. Angular Coherence and Refocusing

Angular coherence plays a large role in our algorithm to establish formulations for both depth estimation and shading constraints. Our goal is to solve for 1) depth map, α^*, and 2) shading in $P = AS$, where P is the central pinhole image of the light-field input L_0, A is the albedo, and S is shading. In order to address the limitations of light-field cameras, we exploit the angular resolution of the data.

Here, we explain why a light-field camera's central pinhole image provides us with an important cue to obtain angular coherence. For an input light-field, L_0, we can shear to refocus the image (introduced by Ng et al. [27]). To shear, the EPI remapping is as follows,

$$\begin{gathered} L_\alpha(x,y,u,v) = L_0(x^f(\alpha), y^f(\alpha), u, v) \\ x^f(\alpha) = x + u(1-\frac{1}{\alpha}) \quad y^f(\alpha) = y + v(1-\frac{1}{\alpha}) \end{gathered} \tag{1}$$

where L_0 is the input light-field image, L_α is the refocused image, (x, y) are the spatial coordinates, and (u, v) are the angular coordinates. The central viewpoint is located at $(u, v) = (0, 0)$.

Given the depth $\alpha^*(x, y)$ for each spatial pixel (x, y), we calculate L_{α^*} by refocusing each spatial pixel to its respective depth. All angular rays converge to the same point on the scene when refocused at α^*, as shown in Fig. 2. We can write this observation as

$$L_{\alpha^*}(x,y,u,v) = L_0(x^f(\alpha^*(x,y)), y^f(\alpha^*(x,y)), u, v) \tag{2}$$

We call this equation the *angular coherence*. Effectively, $L_{\alpha*}$ represents the remapped light-field data of an all-in-focus image. However, utilizing this relationship is difficult because α^* is unknown. From Eqn, 1, the center pinhole image P, where the angular coordinates are at $(u, v) = (0, 0)$, exhibits a unique property: the sheared $x^f(\alpha), y^f(\alpha)$ are independent of (u, v). At every α,

$$L_\alpha(x,y,0,0) = P(x,y) \tag{3}$$

The central angular coordinate always images the same point in the scene, regardless of the focus. This property of refocusing allows us to exploit *photo consistency*, *depth consistency*, and *shading consistency*, shown in Fig. 2. The motivation is to use these properties to formulate depth estimation and shading constraints.

Photo consistency. In L_{α^*}, since all angular rays converge to the same point in the scene at each spatial pixel, the angular pixel colors converge to $P(x, y)$. In high noise scenarios, we use a simple median filter to de-noise $P(x, y)$. Therefore, we represent the photo consistency measure as,

$$L_{\alpha^*}(x,y,u,v) = P(x,y) \tag{4}$$

Depth consistency. Additionally, the angular pixel values should also have the same depth values, which is represented by,

$$\bar{\alpha}^*(x,y,u,v) = \alpha^*(x,y) \tag{5}$$

where $\bar{\alpha}^*$ is just an up-sampled α^* with all angular pixels, (u, v), sharing the same depth for each (x, y). [1]

Shading consistency. Following from the photo consistency of angular pixels for each spatial pixel in L_{α^*}, shading consistency also applies, since shading is viewpoint independent for Lambertian surfaces. Therefore, when solving for shading across all views, *shading consistency* gives us,

$$S(x^f(\alpha^*(x,y)), y^f(\alpha^*(x,y)), u, v) = S(x,y,0,0) \tag{6}$$

3.1. Inverse Mapping

For all three consistencies, the observations only apply to the coordinates in L_{α^*}. To map these observations back to the space of L_0, we need to use the coordinate relationship between L_{α^*} and L_0, as shown in Fig. 2 on the bottom.

$$\begin{gathered} L_0(x^i(\alpha^*), y^i(\alpha^*), u, v) = L_{\alpha^*}(x,y,u,v) \\ x^i(\alpha) = x - u(1-\frac{1}{\alpha}) \quad y^i(\alpha) = y - v(1-\frac{1}{\alpha}) \end{gathered} \tag{7}$$

We use this property to map depth and shading consistency to L_0.

[1] Although depths vary with the viewpoint, (u, v), we can assume the variation of depths between angular pixels is minimal since the aperture is small and our objects are comparatively far away.

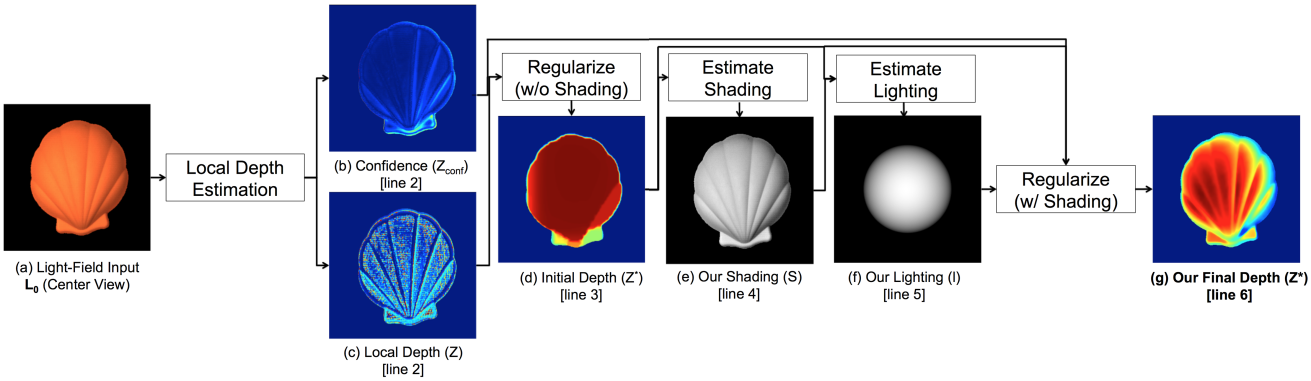


Figure 3. Pipeline. *The pipeline of our algorithm contains multiple steps to estimate the depth of our input light-field image (a). The first is to locally estimate the depth (line 2), which provides us both confidence (b) and local estimation (c). We use these two to regularize depth without shading cues (d) (line 3). The depth is planar, which motivates us to use shading information to refine our depth. We first estimate shading (e) (line 4), which is used to estimate lighting (f) (line 5). We then use the lighting, shading, initial depth, and confidence to regularize into our final depth (g) (line 6).*

Algorithm 1
Depth from Shading, Defocus, and Correspondence

1: **procedure** DEPTH(L_0)
2: Z, Z_{conf} = LocalEstimation(L_0) ▷ Sec. 4.1
3: Z^* = OptimizeDepth(Z, Z_{conf}) ▷ Sec. 4.2
4: S = EstimateShading(L_0) ▷ Sec. 5.1
5: l = EstimateLighting(Z^*, S) ▷ Sec. 5.2
6: Z^* = OptimizeDepth($Z^*, Z_{\text{conf}}, l, S$) ▷ Sec. 5.3
7: **return** Z^*
8: **end procedure**

4. Algorithm

In this section, we discuss local estimation using angular coherence (4.1) and regularization (4.2), corresponding to lines 2 and 3 of the algorithm. Section 5.1 describes shading and lighting estimation and the final optimization. Our algorithm is shown in Algorithm 1 and Fig. 3.

4.1. Depth Cues using Angular Coherence [Line 2]

We start with local depth estimation, where we seek to find the depth α^* for each spatial pixel. We follow Tao et al. [32], which combines defocus and correspondence cues. However, there are some limitations in their approach. Since out-of-focus images still exhibit high contrast, cue responses are incorrect. These situations are common because of lens properties, out-of-focus blur (bokeh), and refocusing artifacts from light-field cameras, as shown in Fig. 4.

We use *photo consistency* (Eq. 4) to formulate an improved metric for defocus and correspondence. From angular coherence (Eq. 2), we want to find α^* such that

$$\alpha^*(x,y) = \underset{\alpha}{\operatorname{argmin}} \, |L_0(x^f(\alpha), y^f(\alpha), u, v) - P(x,y)| \tag{8}$$

The equation enforces all angular pixels of a spatial pixel to equal the center view pixel color, because regardless of α the center pixel color P does not change. We will now reformulate defocus and correspondence to increase robustness of the two measures.

Defocus. Instead of using a spatial contrast measure to find the optimal depth [32], we use Eq. 8 for our defocus measure. The first step is to take the EPI and average across the angular (u, v) pixels,

$$\bar{L_\alpha}(x,y) = \frac{1}{N_{(u,v)}} \sum_{(u',v')} L_\alpha(x,y,u',v') \tag{9}$$

where $N_{(u,v)}$ denotes the number of angular pixels (u, v). Finally, we compute the defocus response by using a measure:

$$D_\alpha(x,y) = \frac{1}{|W_D|} \sum_{(x',y') \in W_D} |\bar{L}_\alpha(x',y') - P(x',y')| \tag{10}$$

where W_D is the window size (to improve robustness). For each pixel in the image, we compare a small neighborhood patch of the refocused image and its respective patch at the same spatial location of the center pinhole image.

Even with refocusing artifacts or high frequency out-of-focus blurs, the measure produces low values for refocusing to non-optimal α. In Fig. 4, we can see that the new measure responses are more robust than responses proposed by Tao et al. [32] (Fig. 4).

Correspondence By applying the same concept as Eqn. 8, we can also formulate a new correspondence measure. To measure photo consistency, instead of measuring the variance of the angular pixels, we measure the difference between the refocused angular pixels at α and their respective center pixel. This is represented by

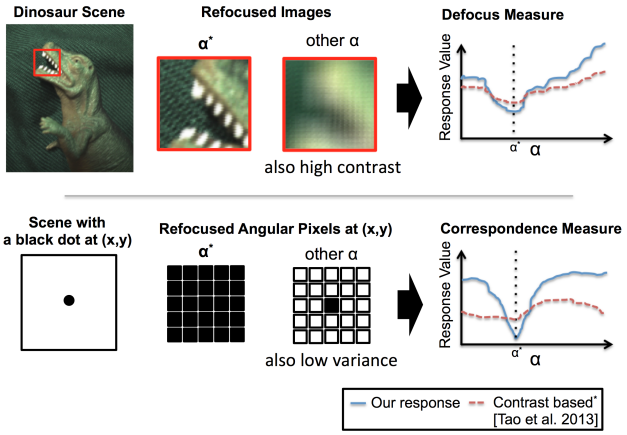


Figure 4. Depth estimation using angular coherence. *On the top, we have a scene with a dinosaur. Even refocused to an unoptimal depth, not equal to α^*, high contrast still exists. By using a contrast based defocus measure, the optimal response is hard to distinguish. On the bottom, we have a scene with a black dot in the center. When refocused at a non-optimal depth, the angular pixels may exhibit the same color as the neighboring pixels. Both the optimal and non-optimal α measures would have low variance. However, by using angular coherence to compute the measures, we can see that, in both cases, the resulting measure better differentiates α^* from the rest, giving us better depth estimation and confidence (also in Fig. 7). Note: For defocus measurement, we inverted the Tao et al. response for clearer visualization.*

$$C_\alpha(x,y) = \frac{1}{N_{(u',v')}} \sum_{(u',v')} |L_\alpha(x,y,u',v') - P(x,y)| \tag{11}$$

Previous work such as Tao et al. [32] only consider the variance in L_α directly, while we also compare to the intended value. This has the following advantages: the measurement is more robust against small angular pixel variations such as noise. See Fig. 4 bottom, where at an incorrect depth, the angular pixels are similar to their neighboring pixels. Measuring the variance will give an incorrect response as opposed to our approach of comparing against the center view.

Confidence and Combining Cues Since the relative confidences of the different cues are important, we surveyed a set of confidence measures. We found Attainable Maximum Likelihood (AML), explained in Hu and Mordohai [19], to be the most effective.

To combine the two responses, for each spatial pixel, we use a simple average of the defocus and correspondence responses weighted by their respective confidences. To find the optimal depth value for each spatial pixel, we use the depth location of the minimum of the combined response curve, which we will label as Z. We used the same AML measure for the new combined response to compute the overall confidence level for local depth estimation, which we label as Z_{conf} (see Fig. 3b,c).

4.2. Regularization w/ Confidence Measure [Line 3]

Up to this point, we have obtained a new local depth estimation. Now the goal is to propagate the local estimation to regions with low confidence.

In our optimization scheme, given Z, the local depth estimation, and its confidence, Z_{conf}, we want to find a new Z^* that minimizes

$$E(Z^*) = \sum_{(x,y)} \lambda_d E_d(x,y) + \lambda_v E_v(x,y) \tag{12}$$

where Z^* is the optimized depth, E_d is our data constraint, and E_v is our smoothness constraint. In our final optimization, we also use E_s, our shading constraint (line 6). In our implementation, we used $\lambda_d = 1$ and $\lambda_v = 4$.

Data constraint (E_d) To weight our data constraint, we want to optimize depth to retain the local depth values with high confidence. Note that since we use light-field data, we have a confidence metric from defocus and correspondence, which may not always be available with other RGBD methods. Therefore, we can establish the data term as follows,

$$E_d(x,y) = Z_{\text{conf}}(x,y) \cdot ||Z^*(x,y) - Z(x,y)||^2 \tag{13}$$

Smoothness constraint (E_v) The smoothness term is the following:

$$E_v(x,y) = \sum_{i=1,2,3} ||(Z^* \otimes F_i)(x,y)||^2 \tag{14}$$

In our implementation, we use three smoothness kernels,

$$F_1 = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix} F_2 = \begin{bmatrix} -1 & 0 & 1 \end{bmatrix} F_3 = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} \tag{15}$$

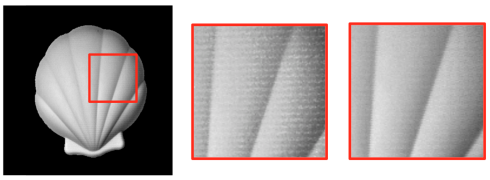
where F_1 is the second derivative and F_2 and F_3 are horizontal and vertical first derivatives respectively.

5. Finding Shading Constraints

The problem with just using the data and smoothness terms is that the smoothness terms do not accurately represent the shape (Fig. 3d). Since smoothness propagates data with high local confidence, depth regularization becomes planar and incorrect (See Fig. 1). Shading information provides important shape where our local depth estimation does not. Before we can add a shading constraint to the regularizer, we need to estimate shading and lighting.

5.1. Shading w/ Angular Coherence [Line 4]

The goal of the shading estimation is to robustly estimate shading with light-field data. We use the decomposition, $P = AS$, where P is the central pinhole image, A



(a) Shading Image (S) (Center View) (b) Without Angular Coherency **(c) With Angular Coherency**

Figure 5. Angular Coherence and Robust Shading. *From the shading image we generate (a), without angular coherency causes noise and unwanted artifacts (b). With angular coherence, the noise reduces. Quantitatively, we can see these effects in Fig. 6.*

is the albedo, and S is the shading. However to improve robustness, we use the full light-field data $L_0 = AS$. Our optimization solves for $S(x, y, u, v)$. In this section, to simplify our notation, we use I to denote L_0, following the standard intrinsic image notation. We use the log space $\log I = \log(A \cdot S)$. We also use $a = i - s$ where the lower case (i, a, s) are the log of (I, A, S) RGB values. We solve for s by using the following error metric,

$$E(s) = \sum_{t=(x,y,u,v)} E_{ls}(t) + E_{la}(t) + E_{ns}(t) + E_{na}(t) + E_{ac}(t). \tag{16}$$

We use a least squares solver to optimize for $s(x, y, u, v)$. To map to $s(x, y)$ (the shading decomposition of P), we take the central viewpoint, $s(x, y, 0, 0)$. We use the shading component of P for lighting and depth refinement for Sec. 5.2 and 5.3.

Depth propagation. Since the shading constraints depend on normals of the entire (x, y, u, v) space, we need to propagate depth and constraints from $Z^*(x, y)$ to $Z^*(x, y, u, v)$. By looking at Fig. 2, we need to map $Z^*(x, y)$ to $\bar{Z}^*(x, y, u, v)$ by using Eqn 6. To map $\bar{Z}^*(x, y, u, v)$ back to the inverse coordinates, we use,

$$Z^*(x^i(\alpha^*), y^i(\alpha^*), u, v) = \bar{Z}^*(x, y, u, v) \tag{17}$$

Local shading and albedo constraint (E_{ls}, E_{la}) To smooth local shading, we look at the 4-neighborhood normals. If the normals are similar, we enforce smoothness.

$$E_{ls}(t) = w_{ls}(t) \cdot ||(s \otimes F_1)(t)||^2$$
$$E_{la}(t) = w_{la}(t) \cdot ||((i - s) \otimes F_1)(t)||^2 \tag{18}$$

where w_{ls} is the average of the dot product between $n(p)$ and w_{la} is the average of the dot product of the RGB chromaticity. F_1 is the second derivative kernel from Eqn. 15.

Nonlocal shading and albedo constraint (E_{ns}, E_{na}) To smooth nonlocal shading, we search for the global closest normals and enforce smoothness. For the pixels with similar normals, we enforce similarity.

$$E_{ns}(t) = \sum_{p,q \in \aleph_{ns}} w_{ns}(p, q) \cdot ||s(p) - s(q)||^2$$
$$E_{na}(t) = \sum_{p,q \in \aleph_{na}} w_{na}(p, q) \cdot ||(i - s)(p) - (i - s)(q)||^2 \tag{19}$$

where p and q represent two unique (x, y, u, v) coordinates within $\aleph_{ns}$ and $\aleph_{na}$, the top 10 pixels with nearest normal and chromaticity respectively. w_{ns} and w_{na} are the dot product between each pairwise normals and chromaticities.

Angular coherence constraint (E_{ac}) So far, we are operating largely similar to shape from shading systems in a single (non light-field) image. We only constrain spatial pixels for the same angular viewpoint. Just like our depth propagation, we can enforce *shading consistency*. We do this by the constraints represented by Eq. 6, as shown in Fig. 2. For each pair of the set of (x, y, u, v) coordinates, we impose the shading constraint as follows,

$$E_{ac}(t) = \sum_{p,q \in \aleph_{ac}} ||s(p) - s(q)||^2 \tag{20}$$

where p, q are the coordinate pairs (x, y, u, v) in $\aleph_{ac}$, all the pixels within the shading constraint. The term plays a large role in keeping our shading estimation robust against typical artifacts and noise associated with light-field cameras. Without the term, the shading estimation becomes noisy and creates errors for depth estimation (Figs. 5, 6).

5.2. Lighting Estimation [Line 5]

With shading, S, we use spherical harmonics to estimate general lighting as proposed by Ramamoorthi and Hanrahan [28] and Basri and Jacobs [5].

$$P = A(x, y) \sum_{k=0}^{8} l_k H_k(Z^*(x, y)) \tag{21}$$

where P is the observed image (L_0), A is the albedo, l are the light source coefficients, and H_k are the spherical harmonics basis functions that take a unit surface normal (n_x, n_y, n_z) derived from $Z^*(x, y)$.

We have computed S. A is estimated as $P = AS$. Therefore, l is the only unknown and can be estimated from these equations using a linear least squares solver.

5.3. Regularization w/ Shading Constraints [Line 6]

With both shading S and lighting l, we can regularize with the shading cue. The new error metric is

$$E(Z^*) = \sum_{(x,y)} \lambda_d E_d(x, y) + \lambda_v E_v(x, y) + \lambda_s E_s(x, y) \tag{22}$$

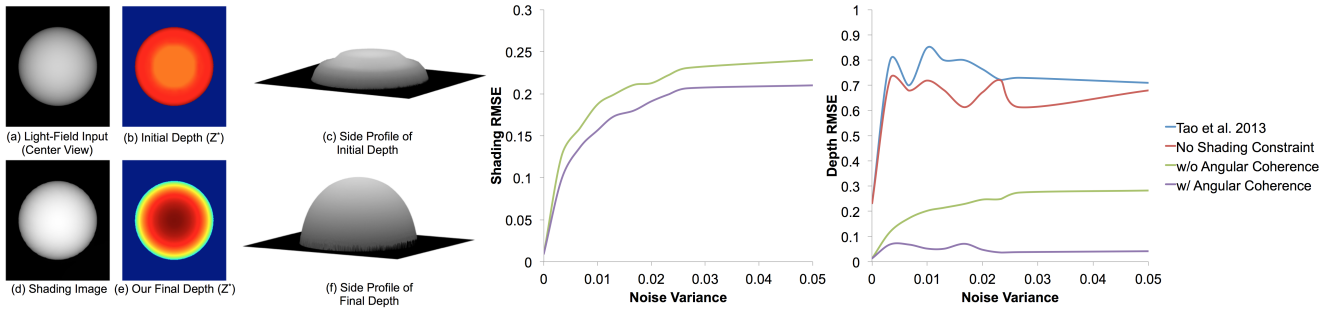


Figure 6. Qualitative and quantitative synthetic measurement. *We have a simple diffuse ball lit by a distant point light-source (a). With just regularization without shading information, our depth estimation does not represent the shape (b,c). With our shading image (d), our depth estimation recovers the ball's surface (e,f). We also added a Gaussian noise with a variable variance. Without the shading constraint, the RMSE against ground truth shading and depth are high. Angular coherence results significantly lower RMSE for both shading and depth.*

where E_d and E_f are the same as Eq. 21 and E_s is our shading constraint. We use $\lambda_s = 2$ in our implementation. We use a non-linear least squares approach with a 8 nearest-neighbors Jacobian pattern to solve for the minimization.

Shading constraint (E_s) To constrain the depth with shading, we want Z^* to satisfy $\sum_{k=0}^{8} l_k H_k(Z^*(x,y)) = S$. Hence, the error term is

$$E_s(x,y) = w_s(x,y) \cdot || \sum_{k=0}^{8} l_k H_k(Z^*(x,y)) - S||^2 \quad (23)$$

where $w_s(x,y) = (1 - Z_{\text{conf}}(x,y))$ to enforce the shading constraint where our local depth estimation is not confident.

6. Results

We validated our algorithm (depth regularized without shading constraints, shading estimation, and depth regularized with shading constraints) using a synthetic light-field image (Fig. 6), and compared our work against other algorithms on real images (Fig. 7). To capture all natural images in the paper, we reverse engineered the Lytro Illum decoder and used varying camera parameters and scenes under different lighting conditions. Please look at our supplementary materials for comprehensive comparisons.

6.1. Synthetic

To validate the depth and shading results of our algorithm, we compare our results to the ground truth depth and shading for a synthetic light-field image of a Lambertian white sphere illuminated by a distant point light source. We added Gaussian noise (zero mean with variance from 0 to 0.03) to the input image. In Fig 6, we see that using shading information helps us better estimate the sphere. With angular coherence constraints on our shading, both depth and shading RMSE are reduced especially with increased noise.

6.2. Natural Images

For natural images, we compare our depth results (depth regularized without and with shading constraints) to the state-of-the-art methods by the Lytro Illum Software, Tao et al. [32], and Wanner et al. [35]; we compare our algorithm to related work by Chen and Koltun [9] as well as Barron and Malik [3] in our supplementary material.

In Fig. 7 top, we have an orange plastic shell, illuminated by an indoor lighting. The Illum produces noisy results. Wanner and Goldlucke regularization propagates errors in regions where local estimation fails. In Tao et al.'s results, graph-cut block artifacts are present. Even without shading constraints, we produce a less noisy result. Our depth estimation recovers the shell shape, including the ridges and curvature. In the middle, we have an image of a cat figurine. Our algorithm recovers the curvature of the body and face. On the bottom, we have an example of a dinosaur toy with varying albedo. The dinosaur teeth, claws, and neck ridges are salient in our results, while other algorithms have trouble recovering these shapes. Using shading gives a significant benefit in recovering the object shapes. Our supplementary materials showcase more comparisons.

7. Conclusion and Future Work

We have proposed a new framework where angular coherence improves the robustness of using cues from defocus, correspondence, and shading. Our optimization framework can be used for consumer grade light-field images to incorporate shading for better shape estimation.

Our algorithm assumes Lambertian surfaces. In future work, more robust approaches could be used for scenes with more varying albedos and occlusions. For example, we could regularize using specularity constraints or reduce effects from specularities by diffuse-specular separation [33]. Additionally, as can be seen in Fig. 6, image noise still corrupts both our depth and shading estimation; more advanced de-noising could be used in the future.

In summary, we have proposed a robust shading-based depth estimation algorithm for light field cameras, suitable for a passive point-and-shoot acquisition from consumer light-field cameras.

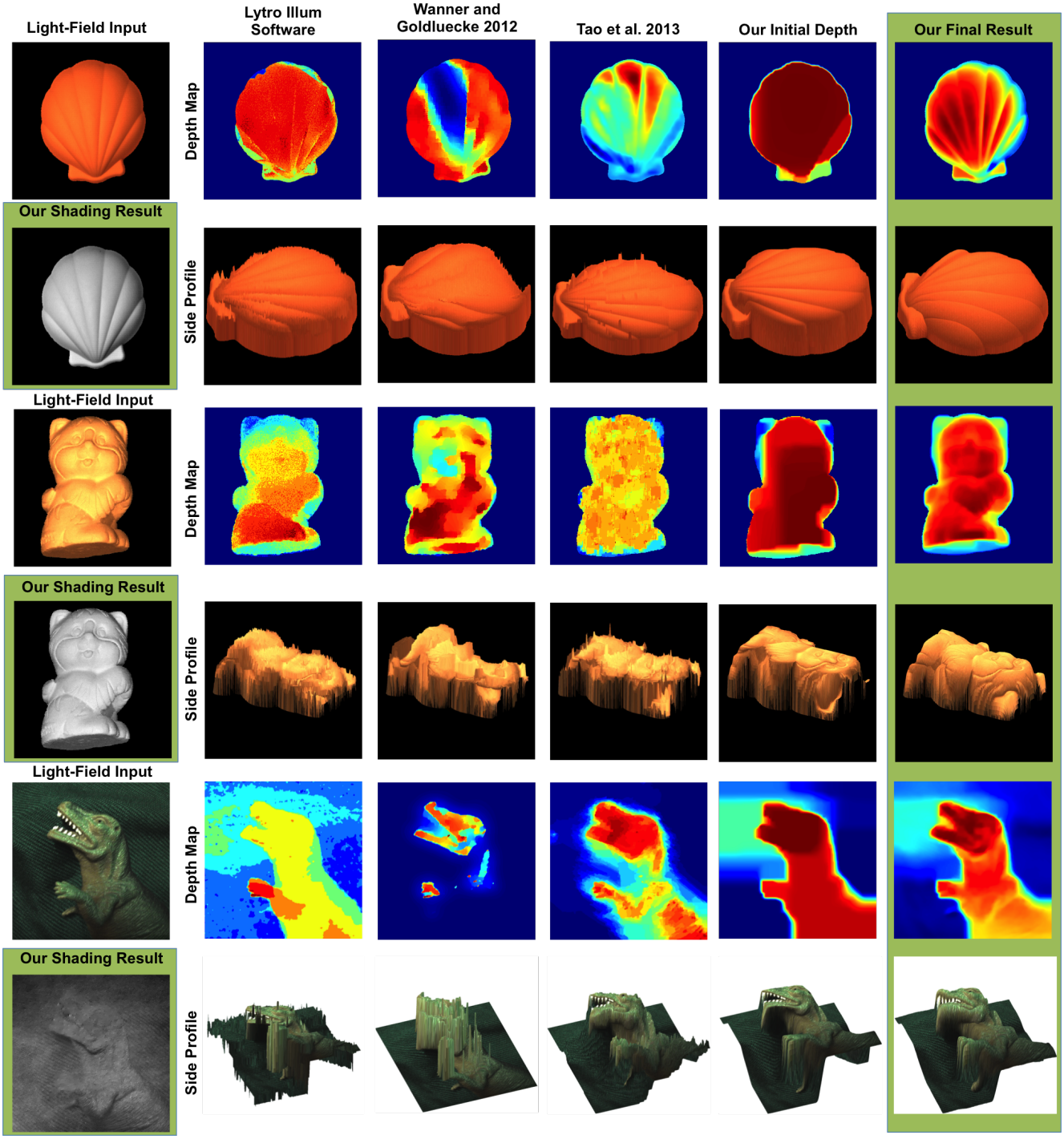


Figure 7. Real World Depth Comparisons.*We compare our work against the light-field depth estimation algorithms: Lytro Illum Software, Wanner and Goldluecke [35], and Tao et al. [32]. On the top, we have a diffuse orange plastic shell, illuminated by a typical indoor area lighting. In our shading estimation, we are able to recover the shape of the shell, including the ridges and curvature. In the middle, we have an image of a cat figurine. We can see that our algorithm is able to recover the curvature of the body and face. On the bottom, we have another example of a dinosaur toy with varying albedo. We can see that the dinosaur teeth, claws, and neck ridges are salient in our results, while other algorithms have trouble recovering these shapes. Moreover, with our initial depth estimation, we already see that our results are smoother and less prone to noise. We can see the shape recovery with the side profiles. We encourage the readers to look through our supplementary materials for more views, examples, and comparisons.*

Acknowledgement

We thank Jong-Chyi Su for generating the synthetic images and comparisons. We acknowledge the financial support from NSF Fellowship DGE 1106400; NSF Grants IIS-1012147 and IIS-1421435; ONR grants N00014-09-1-0741, N00014-14-1-0332, and N00014-15-1-2013; funding from Adobe, Nokia and Samsung (GRO); and support by Sony to the UC San Diego Center for Visual Computing.

References

[1] J. Barron and J. Malik. Color constancy, intrinsic images, and shape estimation. In *ECCV*, 2012.

[2] J. Barron and J. Malik. Shape, albedo, and illumination from a single image of an unknown object. In *CVPR*, 2012.

[3] J. Barron and J. Malik. Intrinsic scene properties form a single rgb-d image. In *CVPR*, 2013.

[4] R. Basri and D. Jacobs. Photometric stereo with general, unknown lighting. In *CVPR*, 2001.

[5] R. Basri and D. Jacobs. Lambertian reflectance and linear subspaces. *PAMI*, 2003.

[6] T. Beeler, D. Bradley, H. Zimmer, and M. Gross. Improved reconstruction of deforming surfaces by canceling ambient occlusion. In *ECCV*, 2010.

[7] A. Bermano, D. Bradley, T. Zund, D. Nowrouzezahrai, I. Baran, O. Sorkine-Hornung, H. Pfister, R. Sumner, B. Bickel, and M. Gross. Facial performance enhancement using dynamic shape space analysis. *ACM Trans. on Graph.*, 2014.

[8] M. K. Chadraker. What camera motion reveals about shape with unknown brdf. In *CVPR*, 2014.

[9] Q. Chen and V. Koltun. A simple model for intrinsic image decomposition with depth cues. In *ICCV*, 2013.

[10] P. Debevec. Rendering synthetic objects into real scenes: bridging traditional and image-based graphics with global illumination and high dynamic range photography. In *SIGGRAPH*, 2012.

[11] K.-D. Durou, M. Falcone, and M. Sagona. Numerical methods for shape-from-shading: A new survey with benchmarks. *Computer Vision and Image Understanding*, 2008.

[12] S. Fanello, C. Keskin, S. Izadi, P. Kohli, and et al. Learning to be a depth camera for close-range human capture and interaction. *ACM Trans. Graph.*, 2014.

[13] B. Freedman, A. Shpunt, M. Machline, and Y. Arieli. Depth mapping using projected patterns. *US Patent*, 2009.

[14] B. Goldleucke and S. Wanner. The variational structure of disparity and regularization of 4d light fields. In *CVPR*, 2013.

[15] S. Gortler, R. Grzeszczuk, R. Szeliski, and M. Cohen. The lumigraph. In *ACM SIGGRAPH*, 1996.

[16] S. Hasinoff and K. Kutulakos. Confocal stereo. *ECCV*, 2006.

[17] C. Hernandez, G. Vogiatzis, and R. Cipolla. Multipleview photometric stereo. *PAMI*, 2008.

[18] B. K. P. Horn. Shape from shading; a method for obtaining the shape of a smooth opaque object form one view. *Ph.D. thesis, Massachusetts Institute of Technology*, 1970.

[19] X. Hu and P. Mordohai. A quantitative evaluation of confidence measures for stereo vision. *PAMI*, 2012.

[20] A. Janoch, S. Karayev, Y. Jia, J. Barron, M. Fritz, K. Saenko, and T. Darrell. A category-level 3D object dataset: putting the kinect to work. In *ICCV*, 2011.

[21] J. Jeon, S. Cho, X. Tong, and S. Lee. Intrinsic image decomposition using structure-texture separation and surface normals. In *ECCV*, 2014.

[22] M. Johnson and E. Adelson. Shape estimation in natural illumination. In *CVPR*, 2011.

[23] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, and M. Gross. Scene reconstruction from high spatio-angular resolution light fields. In *SIGGRAPH*, 2013.

[24] K. Lee and C.-C. Kuo. Shape from shading with a linear triangular element surface model. *IEEE PAMI*, 1993.

[25] M. Levoy and P. Hanrahan. Light field rendering. In *ACM SIGGRAPH*, 1996.

[26] D. Nehab, S. Rusinkiewicz, J. Davis, and R. Ramamoorhti. Color constancy, intrinsic images, and shape estimation. In *ECCV*, 2012.

[27] R. Ng, M. Levoy, M. Bredif, G. Duval, M. Horowitz, and P. Hanrahan. Light field photographhy with a hand-held plenoptic camera. *CSTR 2005-02*, 2005.

[28] R. Ramamoorthi and P. Hanrahan. A signal processing framework for inverse rendering. *ACM Trans. on Graph.*, 2001.

[29] N. Sabater, M. Seifi, V. Drazic, G. Sandri, and P. Perez. Accurate disparity estimation for plenoptic images. In *ECCV LF4CV*, 2014.

[30] S. Seitz and C. Dyer. Photorealistic scene reconstruction by vocal coloring. *IJCV*, 1999.

[31] S. Suwajanakorn, I. Kemelmacher-Shlizerman, and S. Seitz. Total moving face reconstruction. In *ECCV*, 2014.

[32] M. Tao, S. Hadap, J. Malik, and R. Ramamoorthi. Depth from combining defocus and correspondence using light-field cameras. In *ICCV*, 2013.

[33] M. Tao, T.-C. Wang, J. Malik, and R. Ramamoorthi. Depth estimation for glossy surfaces with light-field cameras. In *ECCV LF4CV*, 2014.

[34] A. van Doorn, J. Koenderink, and J. Wagemans. Lightfield and shape from shading. *Journal of Vision*, 2011.

[35] S. Wanner and B. Goldluecke. Globally consistent depth labeling of 4D light fields. In *CVPR*, 2012.

[36] R. J. Woodham. Photometric method for determining surface orientation from multiple images. *Optical Engineering*, 1980.

[37] R. J. Woodham. Gradient and curvature from photometric stereo including local confidence estimation. *Journal of the Optical Society of America*, 1994.

[38] C. Wu, M. Zollhofer, M. Niebner, M. Stamminger, S. Izadi, and C. Theobalt. Real-time shading-based refinement for consumer depth cameras. In *SIGGRAPH Asia*, 2014.

[39] L.-F. Yu, S.-K. Yeung, I.-W. Tai, and S. Lin. Shading-based shape refinement of rgb-d images. In *CVPR*, 2013.

[40] Z. Zhang, P.-S. Tsa, J. Cryer, and M. Shah. Shape from shading: A survey. *PAMI*, 1999.