# High-Frequency Shape and Albedo from Shading using Natural Image Statistics

Jonathan T. Barron and Jitendra Malik
University of California, Berkeley
Berkeley, CA, 94720
{barron, malik}@eecs.berkeley.edu

## Abstract

*We relax the long-held and problematic assumption in shape-from-shading (SFS) that albedo must be uniform or known, and address the problem of "shape and albedo from shading" (SAFS). Using models normally reserved for natural image statistics, we impose "naturalness" priors over the albedo and shape of a scene, which allows us to simultaneously recover the most likely albedo and shape that explain a single image. A simplification of our algorithm solves classic SFS, and our SAFS algorithm can solve the intrinsic image decomposition problem, as it solves a superset of that problem. We present results for SAFS, SFS, and intrinsic image decomposition on real lunar imagery from the Apollo missions, on our own pseudo-synthetic lunar dataset, and on a subset of the MIT Intrinsic Images dataset[15]. Our one unified technique appears to outperform the previous best individual algorithms for all three tasks.*

*Our technique allows a coarse observation of shape (from a laser rangefinder or a stereo algorithm, etc) to be incorporated a priori. We demonstrate that even a small amount of low-frequency information dramatically improves performance, and motivate the usage of shading for high-frequency shape (and albedo) recovery.*

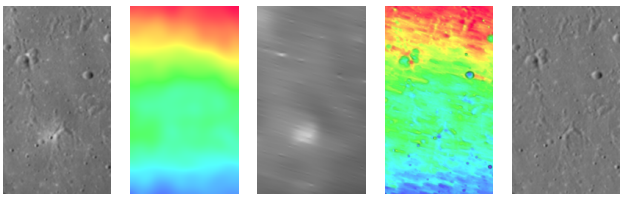
## 1. Introduction

Our work will address what we believe are two of the primary shortcomings of shape from shading:

1. albedo must be uniform and known.
2. shading is a weak cue for low-frequency information.

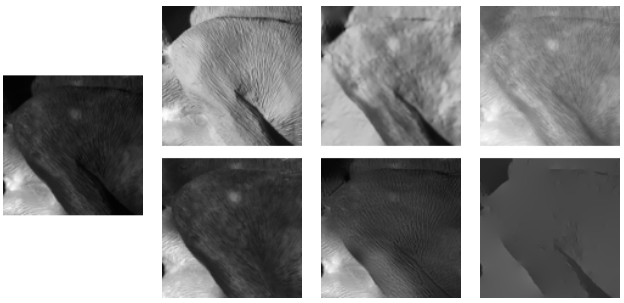
We will ignore the problems of mutual illumination and cast shadows, and we will assume the light source is known.

Determining shape and albedo from a single image is a difficult, under-constrained, and fundamental problem in human and computer vision. A human observer perceives the Mona Lisa neither as paint on a flat canvas nor as a strange shape with uniform albedo, but instead as a 3D woman with light skin and dark hair. This naturally suggests a statistical formulation of the problem, where certain shapes and albedos are more likely than others. We will address the problem of recovering *the most likely albedo and*



(a) Input image (b) Input depth (optional) (c) Output albedo (d) Output depth (e) Output shading

Figure 1. Our SAFS algorithm demonstrated on an image of the moon from Apollo 15. The current best stereo algorithm on this data[6] is used as a low-frequency prior on depth. Our model correctly attributes craters to circular indentations of depth, and maria (dark areas) and ejecta (light areas near craters) to albedo. Depth is visualized with: brightness $\propto \mathbf{n}_z$, color $\propto$ depth (red = near, blue = far).



(a) Input image (b) Ground Truth (c) This paper (d) Retinex[15]

Figure 2. Our SAFS algorithm demonstrated on the intrinsic image decomposition task, using a subset of a test-set image from the MIT intrinsic images dataset. Top row is shading, bottom row is albedo (reflectance). Our algorithm beats Retinex (the previous best algorithm on this dataset) on both error metrics, and produces qualitatively better looking results. Note: our algorithm actually produces a depth-map (not pictured), and our "shading" is a rendering of that depth-map.

*shape that explain a single image*. Using techniques from natural image statistics, we will present novel models for natural albedo images and natural depth maps, whose parameters we learn from training data. Using a novel extension to existing optimization methods we can then produce a depth map and albedo map which maximizes the likelihood of both models.

The second shortcoming of SFS that we will address is that shading is a poor cue for low-frequency (coarse) depth. Shading is directly indicative of only the depth of a point relative to its neighbors: fine-scale variations in depth produce sharp, localized changes in an image, while coarse-scale variations produce very small, subtle changes across an entire image. Bas relief sculptures take advantage of this by conveying the impression of a rich, deep 3D scene, using only the shading produced by a physically shallow object.

We will address this fundamental difficulty by circumventing it, and constructing a framework in which a low-frequency prior on shape (specifically, some low-resolution observation of depth) can be integrated a priori into SAFS and SFS, using a multiscale representation of depth. This framework is complementary to other successes in shape estimation, such as stereo[29, 34], laser range finders, etc, which tend to produce depth maps that are inaccurate at fine scales. Our framework is also compatible with the idea that 3D shape estimation is not entirely bottom-up, but that prior top-down information is used.

We will use multiscale representations of depth and albedo (Laplacian[7] and Gaussian pyramids, respectively), which allow us to integrate low-frequency depth information, and to produce multiscale generalizations of existing natural image statistics models. Our pyramid representation of depth also allows for an effective coarse-to-fine optimization algorithm, based on conjugate gradient descent. We will demonstrate that these benefits of multiscale representations are crucial for accurate reconstruction.

Our work relates to "intrinsic images", which were defined by Barrow and Tenenbaum to include the depth, orientation, reflectance, and illumination of an image[1]. Most work on "intrinsic images" has focused solely on decomposing an image into shading and albedo[15]. We will address the more complete problem of recovering shape (and therefore shading) and albedo, thereby reunifying "intrinsic images" with SFS.

A degenerate case of our SAFS algorithm solves standard SFS while incorporating prior low-frequency information. We present SAFS and SFS results with varied amounts of prior low-frequency information, and show that our SFS algorithm can recover high-frequency shape extremely accurately, while SAFS can recover high-frequency albedo and shape somewhat less accurately, and that accuracy is largely a function of the amount of prior low-frequency information available.

We will demonstrate SAFS and SFS on real and pseudo-synthetic lunar imagery: images from the Apollo 15 mission, and Lambertian renderings of laser scans of Earth's terrain with real albedo maps of the moon. For the Apollo imagery, we use the output of a contemporary stereo algorithm[6] as our low-frequency observation of depth. See Figure 1 for our results on one Apollo image.

We also present results on a subset of the MIT Intrinsic Images dataset[15]. Our algorithm substantially outperforms the grayscale Retinex algorithm, which had previously been the best algorithm for the (single image, grayscale) intrinsic image decomposition task. See Figure 2 for our results on a test-set image.

In Section 2 we review prior work. In Section 3 we formulate our problem as an optimization problem, introduce our priors over albedo and shape, and present our optimization algorithm. In Section 4 we review metrics (and present a novel metric) for evaluating the output of our system and tuning hyperparameters. In Section 5 we introduce a simplification of our model which solves SFS while (optionally) integrating low-frequency depth information. In Section 6 we present results for SAFS and SFS on our real and pseudo-synthetic lunar datasets, and in Section 7 we present results on the MIT intrinsic image dataset. In Section 8 we conclude.

## 2. Prior work

Shape from shading has been the focus of much research since it's formulation by Horn[17]. Many algorithms have been introduced (well surveyed in [5, 38]) and many theoretical results regarding ill-posedness and convergence have been presented. Our rendering formulation is similar to others which optimize over a linearized depth map[31, 35]. Our coarse-to-fine optimization is similar in spirit to multiresolution or multi-grid schemes[8, 18, 31], though our focus is on placing priors on multiscale representations of depth.

The high-frequency nature of shading has been studied in the human and computer vision communities for decades. Blake and Zisserman note that shading is a useful cue for high-frequency depth, and that stereo is better suited for recovering low-frequency depth [4]. Cryer *et al*. demonstrate that integrating the output of SFS and stereo leads to accurate reconstructions where SFS alone fails [9]. Koenderink showed that humans make errors in estimating coarse depth when using only shading[21]. Mamassian *et al*. suggest that contour cues dominate shading cues in the recognition of simple geometric shapes[23]. This view is consistent with ours, in that contours can be used as a low-frequency prior on shape, and shading can provide high-frequency information. That being said, much of our data (see Figure 1) lack geometric contour cues, necessitating the use of shading.

Much work has been done to describe and eliminate

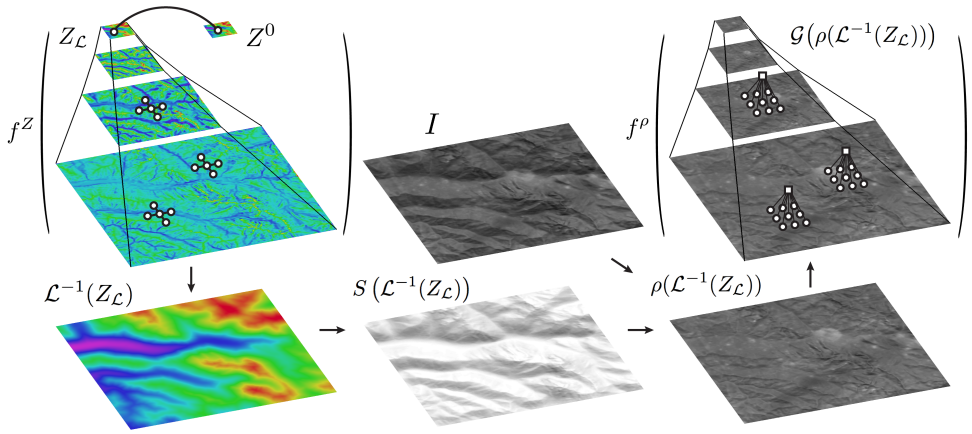



Figure 3. An overview of the calculation of our loss function $f^{safs}(Z_{\mathcal{L}})$. In the upper left we have $f^Z(Z_{\mathcal{L}})$, the negative log-likelihood of a 4-connected MRF over the $Z_{\mathcal{L}}$ (the Laplacian pyramid representation of depth), and a Gaussian prior relating the lowpass residual channel of $Z_{\mathcal{L}}$ with $Z^0$, a coarse observation of depth. In the upper right we have $f^\rho\left(\mathcal{G}\left(\rho(\mathcal{L}^{-1}(Z_{\mathcal{L}}))\right)\right)$, the negative log-likelihood of a multiscale "Field of Experts"-like model over a Gaussian pyramid representation of the albedo implied by $Z_{\mathcal{L}}$. The rest of the figure represents the rendering machinery that allows us to calculate $\rho(\mathcal{L}^{-1}(Z_{\mathcal{L}}))$, the albedo implied by $Z_{\mathcal{L}}$.

the ill-posedness of SFS. To address the "Bas-Relief" ambiguity of Belhumeur *et al*. [2], we assume the direction of light is known. Prados and Faugeras have also shown that concave/convex ambiguities arise unless illumination is attenuated[26], and Forsyth has explored mutual illumination and interreflections[12, 13], all issues which we do not address.

The earliest work in intrinsic images is Land and McCann's Retinex theory[22] (later made practical by Horn[16]), which classifies image gradients as shading or albedo by their magnitude. Later techniques explored Bayesian modeling of image gradients[14], learning classifiers on pyramid coefficients[3], and learning filters to classify image gradients[32, 33]. This line of work is similar to ours in that it involves classifying or modeling local filter output, but all such techniques only produce shading and albedo maps, and all have the severe disadvantage that they do not or cannot reason about underlying shape, while we *explicitly* consider shape.

There is some work on the SAFS problem, all of which must make assumptions to constrain the problem, such as: piece-wise constant albedo (equivalent to Retinex) [24, 28], multiple images [28], or symmetry in shape and albedo[37]. We use a single, non-symmetric image, and rely only on statistical regularities in natural albedo and depth maps.

There is a wealth of work on the statistics of natural images. Huang and Mumford studied the statistics of a variety of representations of natural images[20], and Huang *et al*. found similar trends in range images[19]. We use the former assumption to regularize albedo (as albedo images are a special case of natural images), and the latter assumption to regularize shape. Simoncelli demonstrated descriptive models for adjacent coefficients in wavelet decompositions of images[30], and Portilla et al.[25] demonstrated that Gaussian scale mixtures (GSM) are effective models for denoising such decompositions, an insight which informs our multiscale priors for shape and albedo. Our prior on albedo resembles a multiscale version of the "Field of Experts" [27] with GSM experts[36], an effective and general model for natural image statistics, although we used simple hand-crafted filters rather than learned filters.

## 3. Algorithm

Our problem formulation for "high-frequency shape and albedo from shading" is:

**Input:** image $I$, light direction $L$, [ coarse depth $Z^0$]
**Output:** complete depth $\hat{Z}$, albedo map $\hat{\rho}$

We will use $Z(x,y)$ and $I_{x,y}$ to refer to the depth and intensity of the image at $(x,y)$. $Z^0$ is an optional low-frequency observation of depth. We define $S_{x,y}(Z)$ as the Lambertian rendering of $Z$ at $(x,y)$, illuminated by $L$. Lambertian reflectance states that $I = \rho \cdot S(Z)$. We can therefore define the albedo at $(x,y)$ *implied* by $Z$, $I$ and $L$:

$$\rho_{x,y}(Z) = \frac{I_{x,y}}{S_{x,y}(Z)} \tag{1}$$

Additionally, we define $\mathcal{L}(\cdot)$, which constructs a Laplacian pyramid from an image, $\mathcal{L}^{-1}(\cdot)$, which reconstructs an image from a Laplacian pyramid, and $\mathcal{G}(\cdot)$, which constructs a Gaussian pyramid from an image.

We will construct an optimization problem in which we optimize over $Z_{\mathcal{L}}$, a Laplacian pyramid representation of the depth-map of a scene, to maximize the likelihood of a

prior over $\mathcal{G}\left(\rho\left(\mathcal{L}^{-1}(Z_{\mathcal{L}})\right)\right)$, the Gaussian pyramid representation of the albedo implied by $Z_{\mathcal{L}}$. We will additionally regularize our optimization by placing a prior over $Z_{\mathcal{L}}$. We optimize over the pyramid $Z_{\mathcal{L}}$ rather than the image $Z$ because it allows for descriptive, multiscale priors over depth, and for an effective, multiscale optimization algorithm. Using a pyramid representation of albedo allows us to apply priors to multiple scales, which dramatically improves performance. Our final loss function is:

$$f^{safs}(Z_{\mathcal{L}}) \quad = \quad f^{\rho}\left(\mathcal{G}\left(\rho(\mathcal{L}^{-1}(Z_{\mathcal{L}}))\right)\right) + f^{Z}(Z_{\mathcal{L}}) \quad (2)$$

The calculation of this loss function is illustrated in Figure 3. The output of our algorithm is:

$$\hat{Z} = \mathcal{L}^{-1}\left(\arg\min_{Z_{\mathcal{L}}} f^{safs}(Z_{\mathcal{L}})\right), \quad \hat{\rho} = \rho(\hat{Z}). \qquad (3)$$

In Section 3.1 we formally define $S_{x,y}(Z)$. In Section 3.2 we define $f^{\rho}$ which is a potential on the filter bank response of a multiscale representation of albedo. In Section 3.3 we define $f^{Z}$, which consists of a multiscale MRF and a method of incorporating a low-frequency observation of depth. In Section 3.4 we introduce a novel optimization method for optimizing $f^{safs}$ based on conjugate gradient descent.

### 3.1. Rendering procedure

We will describe our technique for linearizing and then rendering $Z$ with Lambertian reflectance under orthographic projection. This particular linearization is not crucial for SAFS, but it does improve performance. Our technique has two nice advantages over standard linearization: an $(n \times n)$ depth map produces an $(n \times n)$ image as opposed to an $(n-1 \times n-1)$ image, and the normal at $(x, y)$ is a function of $Z$ and all adjacent entries, as opposed to just two.

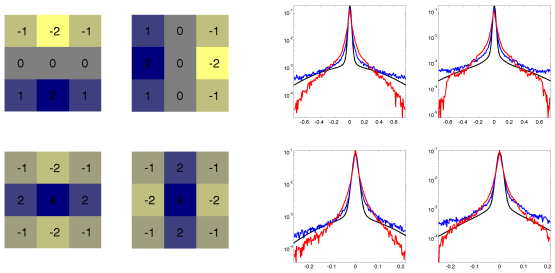
We treat pixel $(x, y)$ as being bounded by four points whose depths we calculate using bilinear interpolation and extrapolation. We then render $S_{x,y}(Z)$ by rendering and then averaging the two triangles formed by those four points. This requires the unit normals of each triangle:

$$\mathbf{n}^{+}_{x,y}(Z) \propto \begin{bmatrix} Z(x-1/2, y-1/2) - Z(x+1/2, y-1/2) \\ Z(x-1/2, y-1/2) - Z(x-1/2, y+1/2) \\ 1 \end{bmatrix}$$

$$\mathbf{n}^{-}_{x,y}(Z) \propto \begin{bmatrix} Z(x-1/2, y+1/2) - Z(x+1/2, y+1/2) \\ Z(x+1/2, y-1/2) - Z(x+1/2, y+1/2) \\ 1 \end{bmatrix}$$

We then render $Z$ with Lambertian reflectance:

$$S_{x,y}(Z) = \tfrac{1}{2}\left(\max(0, L \cdot \mathbf{n}^{+}_{x,y}(Z)) + \max(0, L \cdot \mathbf{n}^{-}_{x,y}(Z))\right). \qquad (4)$$



(a) Filter bank (b) Filter responses

Figure 4. On the left, we have the filter bank we use for $\mathcal{G}\left(\rho(\mathcal{L}^{-1}(Z_{\mathcal{L}}))\right)$, the Gaussian pyramid representation of albedo. On the right, we have log-histograms of the responses for the filter bank on test-set albedos (blue), on test-set images (red), and the GSM model we learn (black). Note that the distribution of filter response on albedo images is different than that of natural images (shading $\times$ albedo) — albedo is much more kurtotic and heavy-tailed. This difference is crucial to the success of our technique.

### 3.2. Priors over albedo

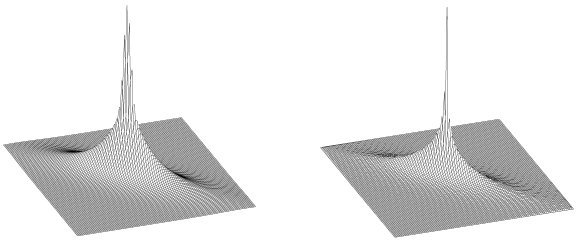
As a model for albedo, we will construct a filter bank (Figure 4(a)) of oriented edge and bar filters, and place Gaussian Scale Mixture[25] (GSM) potentials on the filter responses. This model resembles the "Field of Experts" model (FoE) in [36], but we do not learn our filter bank. We experimented with many sets of hand-designed and learned filters (using the basis rotation technique of [36]), but this simple $3 \times 3$ oriented bar and edge filter bank worked best. The GSM potentials are learned from training albedo data using Expectation Maximization.

Unlike other FoE-like models, we apply our model to a Gaussian pyramid rather than an image. This is equivalent to having an expanded, multi-scale filter bank. We do this because, as opposed to tasks such as image denoising where the image is only corrupted at the finest scale, in SAFS low-frequency errors in the depth-map may corrupt the implied albedo map at low-frequency scales.

We will use the negative log-likelihood of this multi-scale FoE-like model as $f^{\rho}(\rho_{\mathcal{G}})$ in Eq. 2 (where $\rho_{\mathcal{G}} = \mathcal{G}\left(\rho(\mathcal{L}^{-1}(Z_{\mathcal{L}}))\right)$, the Gaussian pyramid representation of the albedo implied by $Z_{\mathcal{L}}$):

$$f^{\rho}(\rho_{\mathcal{G}}) = -\sum_{k=1}^{K} \lambda^{\rho}_{k} \sum_{c \in \mathcal{C}^{\rho}_{k}} \sum_{i=1}^{4} \log\left(\sum_{j=1}^{M} \alpha_{ijk} \cdot \mathcal{N}\left(\mathbf{J}_i^{\mathrm{T}} \mathbf{x}_{(c)}; \mu_{ik}, \sigma^2_{ijk}\right)\right) \qquad (5)$$

Where $K$ is the number of levels of the Gaussian pyramid, $\mathcal{C}^{\rho}_{k}$ are the maximal cliques ($3 \times 3$ patches) of the $k$'th level of $\rho_{\mathcal{G}}$, $\mathbf{x}_{(c)}$ is the $3 \times 3$ patch in $\rho_{\mathcal{G}}$ corresponding to clique $c$, and $\mathbf{J}_i$ is the $i$'th filter, of which there are 4. Regarding each GSM, $\alpha_{ijk}$ are the mixing weights of Gaussian $ik$ (of which there are $M = 50$), each of which has variance $\sigma^2_{ijk}$, and all



(a) A bivariate GSM model (b) the data used to train that model

Figure 5. The bivariate GSM used in one scale of our MRF model over adjacent coefficients in $Z_{\mathcal{L}}$ (the Laplacian pyramid representations of depth), and the data used to train that model.

of which have mean $\mu_{ik}$. $\lambda_k^{\rho}$ are the hyperparameters that weight each scale of the prior, and are tuned to maximize SAFS performance on the training set.

Our filter bank and the GSM potentials for the finest scale of $\rho_{\mathcal{G}}$ on the albedo maps in the MIT Intrinsic Images dataset is shown in Figure 4. Consider the insight of [36] concerning filter selection for image denoising: that filters should fire rarely on natural images but frequently on all other images. For our task, this insight translates to: filters should fire infrequently on albedo images, but frequently on natural images (albedo $\times$ shape). Our filter bank has this property, as evidenced by the highly kurtotic distribution of albedo compared to the less kurtotic distribution of natural images. Because variations in $Z$ tend to be bumps and creases, and because bumps and creates in shapes create bars and edges in the resulting inferred albedo $\rho(Z)$, these filters tend to fire on (and therefore penalize) shape-like variations in inferred albedo.

### 3.3. Priors over shape

We have two goals when constructing priors over the Laplacian pyramid representation of depth $Z_{\mathcal{L}}$: 1) we would like the residual low-pass level of the pyramid to stay close to our low-pass observation $Z^0$ (which is assumed to be the same size as $Z_{\mathcal{L}}[K]$, the top level of $Z_{\mathcal{L}}$), and 2) we would like to regularize $Z_{\mathcal{L}}$ using a statistical model learned from example depth maps. The first goal is accomplished by assuming that $Z^0$ is a noisy observation of $Z_{\mathcal{L}}[K]$ (where noise is Gaussian and i.i.d.) and the second goal is accomplished by maximizing the log-likelihood of $Z_{\mathcal{L}}$ under a 4-connected multiscale MRF, in which each edge potential is a bivariate Gaussian Scale Mixture.

We will use the negative log-likelihood of these priors as $f^Z(Z_{\mathcal{L}})$ in Eq. 2:

$$f^Z(Z_{\mathcal{L}}) = \lambda_K^Z \|Z_{\mathcal{L}}[K] - Z^0\|_2^2 - \sum_{k=1}^{K-1} \lambda_k^Z \sum_{c \in \mathcal{C}_k} \log \left( \sum_{j=1}^{M} \alpha_{jk} \cdot \mathcal{N}\left(\mathbf{x}_{(c)}; \mu_k, s_{jk} \cdot \Sigma_k\right) \right) \quad (6)$$

This is similar to Eq. 5, but we have $K-1$ bivariate GSMs (each with a single covariance matrix $\Sigma_k$) instead of filter banks, and a squared-error term against $Z^0$ at level $K$. $\lambda_k^Z$ are the hyperparameters for each level, and are tuned to maximize SAFS performance on the training set.

One bivariate GSM model is visualized in Figure 5, with the training data used to learn that model. Our model captures the correlation of adjacent coefficients and the heavy-tailed, kurtotic nature of the distribution.

### 3.4. Optimization

We will optimize over $Z_{\mathcal{L}}$ using nonlinear conjugate gradient descent (CG). The properties of our problem require two modifications to standard CG:

It is possible to construct multiple Laplacian pyramids that reconstruct into identical images. For example, one could construct a pyramid in which the entire image is in the high-pass channel of the pyramid. This is a problem, as our priors assume that $Z_{\mathcal{L}}$ is "valid" — that all signal of a certain frequency is contained only in a certain level of the pyramid. We therefore require a guarantee that, in optimization, $Z_{\mathcal{L}}$ is always valid.

An "invalid" pyramid $Z_{\mathcal{L}}$ can be made valid by reconstructing an image, and then constructing a pyramid from that reconstructed image (both using standard methods[7]):

$$\mathcal{V}(Z_{\mathcal{L}}) = \mathcal{L}\left(\mathcal{L}^{-1}(Z_{\mathcal{L}})\right). \quad (7)$$

It can be demonstrated that $\mathcal{V}(x)$ is a linear system, and that if $Z_{\mathcal{L}}$ is valid, $\mathcal{V}(Z_{\mathcal{L}}) = Z_{\mathcal{L}}$. Therefore, given some valid Laplacian pyramid $Z_{\mathcal{L}}$, some vector $x$ of equal size, and some scalar $\alpha$, $\mathcal{V}(Z_{\mathcal{L}} + \alpha x) = Z_{\mathcal{L}} + \alpha \mathcal{V}(x)$. Therefore, using $\mathcal{V}(\Delta_{Z_{\mathcal{L}}})$ in place of $\Delta_{Z_{\mathcal{L}}}$ in CG ensures that $Z_{\mathcal{L}}$ is always in the space of valid Laplacian pyramids, as the conjugate direction is a linear combination of gradients.

Optimizing in the space of valid Laplacian pyramids enables our second modification to CG, in which we do coarse-to-fine optimization. This prevents coarse-scale image features from being wrongly attributed to fine-scale shape features. We iterate over the $K$ levels of $L_z$ from $K$ to 1, and at each iteration we optimize over levels $k$ through $K$ until convergence. We optimize over levels $k$ through $K$ while still guaranteeing a valid pyramid by setting $\Delta_{Z_{\mathcal{L}}}$ at levels 1 through $k-1$ to 0 before calculating $\mathcal{V}(\Delta_{Z_{\mathcal{L}}})$.

Pseudocode for our modified CG can be found in Algorithm 1. Changes to standard CG are indicated with color. An animation of $Z$ and $\rho$ during optimization can be found at http://www.eecs.berkeley.edu/~barron/.

Regarding implementation, efficient CG requires efficient calculation of the $f^{safs}(Z_{\mathcal{L}})$ of $\nabla_{Z_{\mathcal{L}}} f^{safs}(Z_{\mathcal{L}})$. Calculating $f^{safs}(Z_{\mathcal{L}})$ is straightforward, but calculating $\nabla_{Z_{\mathcal{L}}} f^{\rho}\left(\mathcal{G}\left(\rho(\mathcal{L}^{-1}(Z_{\mathcal{L}}))\right)\right)$ requires these non-obvious properties of Laplacian and Gaussian pyramids: $\nabla f(\mathcal{L}(\cdot)) = \mathcal{G}(\nabla f(\cdot))$, and $\nabla f(\cdot) = \mathcal{L}^{-1}(\nabla f(\mathcal{G}(\cdot)))$.

**Algorithm 1** Laplacian Pyramid Conjugate Gradient. Colors indicate differences from standard nonlinear CG (Blue = coarse-to-fine, red = valid Laplacian pyramid)

$x_0 \leftarrow \mathcal{L}(Z^0)$
**for** $k = K$ to 1 **do**
  $\Lambda_0 \leftarrow -\nabla_x f(x_0)$
  **repeat**
    $\Delta_n \leftarrow -\nabla_x f(x_n)$
    $\Delta_n[\text{level} < k] \leftarrow 0$
    $\Delta_n \leftarrow \mathcal{L}^{-1}(\mathcal{L}(\Delta_n))$
    **compute** $\beta_n$ // using Polack-Ribiere
    $\Lambda_n \leftarrow \Delta_n + \beta_n \Lambda_{n-1}$
    $\alpha_n \leftarrow \arg\min_{\alpha_n} f(x_n + \alpha_n \Lambda_n)$ // linesearch
    $x_{n+1} \leftarrow x_n + \alpha_n \Lambda_n$
  **until** converged
**end for**

## 4. Evaluation

Choosing a good error metric is non-trivial, and has been discussed at great length [9, 18, 31]. We choose our error metrics for SAFS and SFS under two goals: accurate shape, and accurate appearance under different illuminations. For the first goal we use mean-squared error (MSE) between $Z$ and the true $Z^*$:

$$Z\text{-MSE}(Z, Z^*) = \frac{1}{n} \sum_{x,y} (Z(x,y) - Z^*(x,y))^2 \,. \tag{8}$$

$Z$-MSE alone does not address our second goal, as very small errors in $Z$ can create large errors in $S(Z)$ and $\rho(Z)$. We considered using the error between the true image and our predicted image, the gradient error [31], or the error in the recovered albedo, but none of these directly satisfy our goal that $Z$ "look good" when rendered under different illuminations. We therefore propose a novel metric that is an approximation of the total MSE between all renderings of $Z$ and $\rho$, and $Z^*$ and $\rho^*$, under all lighting conditions.

First, we will define the effective normal at pixel $(x, y)$:

$$\mathbf{n}_{x,y}(Z) = \frac{1}{2}(\mathbf{n}^+_{x,y}(Z) + \mathbf{n}^-_{x,y}(Z)) \,. \tag{9}$$

After omitting $\max(0, \cdot)$ from Lambertian reflectance, we can define our error metric at a single pixel:

$$\int \left(\rho_{x,y}(L \cdot \mathbf{n}_{x,y}(Z)) - \rho^*_{x,y}(L \cdot \mathbf{n}_{x,y}(Z^*))\right)^2 dL \,. \tag{10}$$

We define $\mathbf{v} = \rho_{x,y}\mathbf{n}_{x,y}(Z) - \rho^*_{x,y}\mathbf{n}_{x,y}(Z^*)$. Integrating over $L$ on only the camera-facing hemisphere ($L_3 \geq 0$), Eq. 10 reduces to:

$$\frac{\pi^2}{4}\left((\mathbf{v}_1 + \mathbf{v}_2)^2 + 2\mathbf{v}_3^2\right) \,. \tag{11}$$

We define $I$-MSE$(Z, Z^*)$ as the mean of Eq. 11 over all $(x, y)$, and use this as our error metric for visual fidelity.

When tuning the hyperparameters ($\lambda$) of our model, we evenly minimize both error metrics by minimizing:

$$\frac{\sum Z\text{-MSE}(\hat{Z}, Z^*)}{\sum Z\text{-MSE}(Z^0, Z^*)} + \frac{\sum I\text{-MSE}(\hat{Z}, Z^*)}{\sum I\text{-MSE}(Z^0, Z^*)} \,. \tag{12}$$

Where the summations are over the images in our training set.

## 5. Shape from shading

Our SAFS problem formulation can be reduced to one that solves SFS, while allowing low-frequency depth information to be integrated. The problem formulation for "high-frequency shape from shading" is:

**Input:** image $I$, light direction $L$, [ coarse depth $Z^0$]
**Output:** complete depth $\hat{Z}$

Instead of placing a prior over implied albedo, here we will simply minimize the squared error between $\mathcal{G}\big(S(\mathcal{L}^{-1}(Z_\mathcal{L}))\big)$ and $\mathcal{G}(I)$ (weighted appropriately at each scale). This can be interpreted as a special case of SAFS in which albedo is assumed to be 1, or as a multiscale generalization — in both $Z$ and $I$ — of other SFS methods based on linearizing a depth-map[31, 35]). Our loss function is:

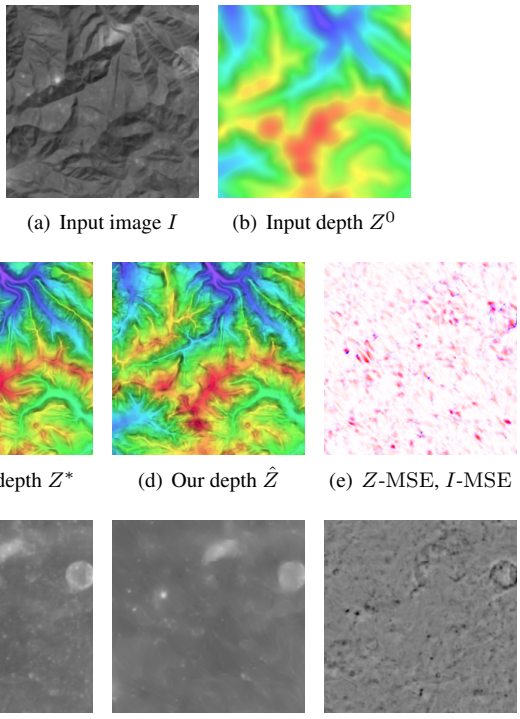
$$f^{sfs}(Z_\mathcal{L}) = \sum_{k=1}^{K} \lambda^I_k \|\mathcal{G}\big(S(\mathcal{L}^{-1}(Z_\mathcal{L}))\big)[k] - \mathcal{G}(I)[k]\|_2^2 + f^Z(Z_\mathcal{L}) \,. \tag{13}$$

## 6. Results: Lunar SAFS

We will first demonstrate our SAFS and SFS algorithms on real and pseudo-synthetic lunar imagery. Our real lunar data are images from the Apollo 15 mission (for which we have no ground-truth) with a contemporary stereo algorithm[6] providing our low-frequency depth $Z^0$. Our pseudo-synthetic lunar dataset is laser scans of the terrain near Puget Sound[1], and lunar albedo maps from the Clementine mission[2], rendered under orthographic projection with Lambertian reflectance. In SFS we use the same data with $\rho = 1$. The dataset is 40 images (20 training, 20 test), each 288×288 pixels (we only use the innermost 256×256 pixels when evaluating error or visualizing). In the high-frequency case, $K = 4$, and $Z^0$ is $Z^*$ downsampled by a factor of $1/64$ in area, with added Gaussian noise ($\sigma = 1$). We train our priors over albedo and shape on the entire training set, but for efficiency we use only 4 training images when optimizing over the $\lambda$ hyperparameters.

As baselines for SAFS, we will use two methods to separate our images into shading and albedo components (the

[1] www.cc.gatech.edu/projects/large_models/ps.html
[2] nssdc.gsfc.nasa.gov/planetary/clementine.html



(a) Input image $I$ (b) Input depth $Z^0$

(c) True depth $Z^*$ (d) Our depth $\hat{Z}$ (e) $Z$-MSE, $I$-MSE

(f) True albedo $\rho(Z^*)$ (g) Our albedo $\rho(\hat{Z})$ (h) $3\times(\rho(\hat{Z})$ - $\rho(Z^*))$

Figure 6. High Frequency ($K = 4$) SAFS output on a test-set image. Depth is visualized as in Figure 1. Error in 6(e) is visualized with (Red $\propto$ $Z$-MSE$(\hat{Z}, Z^*)$, Blue $\propto$ $I$-MSE$(\hat{Z}, Z^*)$).

Table 1. SAFS Results

| Algorithm | **High Frequency** | | Complete | |
|---|---|---|---|---|
| | $Z$-MSE | $I$-MSE | $Z$-MSE | $I$-MSE |
| Retinex[15] + SFS | 131% 2 | 51% 0.02 | 607% 364 | 103% 0.1 |
| Tappen[33] + SFS | 326% 4 | 242% 0.1 | 641% 384 | 230% 0.2 |
| This paper (single scale) | 25% 0.3 | 37% 0.02 | 94% 56 | 66% 0.07 |
| This paper (no Z smoothness) | 15% 0.2 | 30% 0.01 | 93% 56 | 75% 0.08 |
| This paper (all priors) | **5.4% 0.06** | **9.8% 0.005** | **92% 55** | **62% 0.06** |

Table 2. SFS Results

| Algorithm | **High Frequency** | | Complete | |
|---|---|---|---|---|
| | $Z$-MSE | $I$-MSE | $Z$-MSE | $I$-MSE |
| Polynomial SFS[10] | 10% 0.1 | 37% 0.07 | 61% 37 | 30% 0.2 |
| Falcone Sagona[11] | 54% 0.7 | 120% 0.2 | 94% 56 | 113% 0.6 |
| This paper (single scale) | 41% 0.5 | 77% 0.2 | 74% 44 | 76% 0.4 |
| This paper (no Z smoothness) | 3.2% 0.04 | 3.6% 0.007 | 58% 35 | 33% 0.2 |
| This paper (all priors) | **0.5% 0.007** | **1% 0.002** | **45% 27** | **29% 0.2** |

SAFS and SFS results with and without prior low-frequency information. "High Frequency" correspond $K = 4$, where $Z^0$ was created by downsampling $Z^*$ by a factor of $1/64$ in area. In the "Complete" case, $Z^0 = 0$, $K = 7$, and $Z$-MSE is shift-invariant. Errors are shown relative to $Z^0$ in percent, and in absolute terms.

Table 3. Intrinsic Image Results

| Algorithm | LMSE | $I$-MSE |
|---|---|---|
| GR-Retinex[15] + SFS | 0.118 | 2.205 |
| This paper | **0.079** | **0.791** |

Performance on our test-set of the MIT intrinsic images dataset[15]. We report LMSE, the error metric proposed by [15], and our $I$-MSE error metric (note that LMSE is not dependent on SFS). We outperform the grayscale Retinex algorithm (the previous best algorithm for this dataset) on both error metrics by a large margin.

grayscale Retinex implementation of [15], and Tappen *et al.*'s technique in [33]), and we then use our SFS algorithm (the best performing algorithm on this dataset) on the recovered shading. As baselines for SFS, we will use Falcone and Sagona's method[11] and Polynomial SFS[10]. For both tasks, we also present two degenerate cases of our algorithm: a "single scale" model ($K = 1$) with smoothness priors over $Z$, and a multiscale model with no smoothness priors over $Z$. The "single scale" model is effectively an improved version of Tsai and Shah's method[35], which produces results that are comparable to many other SFS techniques[38] (we tried Tsai and Shah's method itself, but our baseline outperformed it by a large margin). In all non-multiscale baselines, we incorporate $Z^0$ by using it as an initialization for $L_z$, and in post-processing by using Laplacian pyramids to replace the $K$'th level of $\hat{Z}$ with $Z^0$ (similar to [9]).

Tables 1 and 2 show results for SAFS and SFS with different amounts of prior low-frequency depth. Many baseline techniques produce worse depth maps than they began with. Our single-scale algorithm dramatically underperforms the multiscale algorithm, demonstrating that multiscale representation and optimization is more effective than regularizing a flat representation. Multiscale smoothness priors over $Z$ do improve performance, though the improvement is most significant in the "high frequency" case.

When the prior low-frequency information is removed in the "Complete" case, our algorithm's advantage over the baselines lessens. In terms of absolute error *all algorithms perform poorly*. This reflects the difficulty in using shading to estimate low-frequency depth (and therefore albedo), and validates our "high frequency" problem formulation. Still, our model consistently outperforms all baselines even when given no low-frequency information, especially in SAFS.

## 7. Results: Intrinsic image decomposition

We evaluate our SAFS algorithm on the intrinsic images task using the MIT Intrinsic Images dataset[15]. We compare against grayscale Retinex, the current best algorithm on this dataset ([32] performs slightly better).

Comparing our algorithm to Retinex is difficult, as we assume the light direction is known, and that ground-truth shape is available for training. This necessitated the use of photometric stereo (each object in the dataset was imaged under multiple light directions) to estimate ground-truth depth maps for each object and the illumination for each image (modeled with a directional light and an ambient term). These ground-truth depths allow us to train priors on $Z$, and allow us to calculate $I$-MSE. Ground truth depth could also be used to produce low-frequency observations

of depth, but this would give us an unfair advantage. Only the estimated light directions are used as input by our SAFS algorithm.

We selected 5 training images (cup1, panther, paper1, squirrel, teabag2) and 5 test images (cup2, deer, paper2, raccoon, teabag1) from the dataset. We selected only regions of the images that did not contain object boundaries or shadowed regions, which our algorithm cannot handle. In Table 3 we present results for grayscale Retinex, and our SAFS algorithm. We beat the grayscale Retinex algorithm (the previous best algorithm for this dataset) by a wide margin: 33% in LMSE, the error metric of [15], and 64% in $I$-MSE. LMSE considers only albedo and shading, while $I$-MSE considers albedo and shape, and therefore shading. An example of our algorithm compared to Retinex can be see in Figure 2.

## 8. Conclusion

We have addressed two of the fundamental issues that limit shape-from-shading: the assumption of a uniform or known albedo, and the difficulty in estimating low-frequency shape. We have presented an algorithm that can recover shape and albedo from a single image better than any previous algorithm, on a challenging dataset that we have presented and on a subset of the MIT Intrinsic Images dataset. Our algorithm can incorporate low-frequency priors on shape, and we have shown that accuracy in SAFS and SFS is largely dependent on such low-frequency information. A simplification of our algorithm outperforms classic SFS algorithms, especially when given low-frequency information.

Our technique depends entirely on our novel assumption of a natural albedo and *shape*, which appears to be much more effective than past algorithms [3, 15, 32, 33] which assume properties of albedo and *shading*, and never consider shape. This difference allows us to outperform the previous best intrinsic image algorithms.

We have demonstrated that by unifying SFS and intrinsic image decomposition into the more comprehensive problem of SAFS, we produce significantly better results than if we address either problem independently.

**Acknowledgements:** Thanks to Ara Nefian, Michael Broxton, Ady Ecker, Stefan Roth, Rob Fergus, and David Forsyth. This work was supported by the NSF GRF Fellowship and ONR MURI N00014-06-1-0734.

## References

[1] H. Barrow and J. Tenenbaum. Recovering intrinsic scene characteristics from images. In *Computer Vision Systems*, pages 3–26. Academic Press, 1978. 2522

[2] P. Belhumeur, D. Kriegman, and A. Yuille. The Bas-Relief Ambiguity. *IJCV*, 35(1):33–44, 1999. 2523

[3] M. Bell and W. T. Freeman. Learning local evidence for shading and reflectance. *ICCV*, 2001. 2523, 2528

[4] A. Blake, A. Zisserman, and G. Knowles. Surface descriptions from stereo and shading. *IVC*, 3(4):183–191, 1986. 2522

[5] M. J. Brooks and B. K. P. Horn. *Shape from shading*. MIT Press, 1989. 2522

[6] M. Broxton, A. V. Nefian, Z. Moratto, T. Kim, M. Lundy, and A. V. Segal. 3d lunar terrain reconstruction from apollo images. In *ISVC (1)*, pages 710–719, 2009. 2521, 2522, 2526

[7] P. J. Burt and E. H. Adelson. The laplacian pyramid as a compact image code. *IEEE Trans. on Comm.*, 31:532–540, 1983. 2522, 2525

[8] A. Crouzil, X. Descombes, and J.-D. Durou. A multiresolution approach for shape from shading coupling deterministic and stochastic optimization. *PAMI*, 25(11):1416 – 1421, 2003. 2522

[9] J. E. Cryer, P. sing Tsai, and M. Shah. Combining shape from shading and stereo using human vision model. Technical report, UCF, 1992. 2522, 2526, 2527

[10] A. Ecker and A. D. Jepson. Polynomial shape from shading. *CVPR*, 2010. 2527

[11] M. Falcone and M. Sagona. An algorithm for the global solution of the shape-from-shading model. *ICIAP*, 1997. 2527

[12] D. Forsyth and A. Zisserman. Reflections on shading. *TPAMI*, 13(7):671–679, 1991. 2523

[13] D. A. Forsyth. Variable-source shading analysis. *Int. J. Comput. Vision*, 91:280–302, February 2011. 2523

[14] W. T. Freeman and P. A. Viola. Bayesian model of surface perception. In *NIPS*. MIT Press, 1997. 2523

[15] R. Grosse, M. K. Johnson, E. H. Adelson, and W. T. Freeman. Ground-truth dataset and baseline evaluations for intrinsic image algorithms. *ICCV*, 2009. 2521, 2522, 2527, 2528

[16] B. K. P. Horn. Determining lightness from an image. *Computer Graphics and Image Processing*, 3(1):277–299, 1974. 2523

[17] B. K. P. Horn. Obtaining shape from shading information. In *The Psychology of Computer Vision*, pages 115–155, 1975. 2522

[18] B. K. P. Horn. Height and gradient from shading. *IJCV*, 5(1):37–75, 1990. 2522, 2526

[19] J. Huang, A. B. Lee, and D. Mumford. Statistics of range images. *CVPR*, 2000. 2523

[20] J. Huang and D. Mumford. Statistics of natural images and models. *CVPR*, 1999. 2523

[21] J. Koenderink, A. Van Doorn, C. Christou, and J. Lappin. Perturbation study of shading in pictures. *Perception*, 25(9):1009–1026, 1996. 2522

[22] E. H. Land and J. J. McCann. Lightness and retinex theory. *J. Opt. Soc. Am.*, 61(1):1–11, 1971. 2523

[23] P. Mamassian, D. Kersten, and D. C. Knill. Categorical local-shape perception. *Perception*, 25:95–107, 1996. 2522

[24] A. Ortiz and G. Oliver. Shape from shading for multiple albedo images. *ICPR*, 1:786 –789, 2000. 2523

[25] J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli. Image denoising using scale mixtures of gaussians in the wavelet domain. *IEEE Trans. Image Process*, 12:1338–1351, 2003. 2523, 2524

[26] E. Prados and O. Faugeras. Shape from shading: A well-posed problem? *CVPR*, 2005. 2523

[27] S. Roth and M. J. Black. Fields of experts: A framework for learning image priors. *CVPR*, 2005. 2523

[28] D. Samaras, D. Metaxas, P. Fua, and Y. G. Leclerc. Variable albedo surface reconstruction from stereo and shape from shading. *CVPR*, 2000. 2523

[29] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 2002. 2522

[30] E. P. Simoncelli. Statistical models for images: Compression, restoration and synthesis. In *In 31st Asilomar Conf on Signals, Systems and Computers*, pages 673–678, 1997. 2523

[31] R. Szeliski. Fast shape from shading. *CVGIP: Image Underst.*, 53(2):129–153, 1991. 2522, 2526

[32] M. Tappen, E. Adelson, and W. Freeman. Estimating intrinsic component images using non-linear regression. *CVPR*, 2006. 2523, 2527, 2528

[33] M. F. Tappen, W. T. Freeman, and E. H. Adelson. Recovering intrinsic images from a single image. *TPAMI*, 27(9):1459–1472, 2005. 2523, 2527, 2528

[34] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon. Bundle adjustment - a modern synthesis. *ICCV*, 1999. 2522

[35] P.-S. Tsai and M. Shah. Shape from shading using linear approximation. *IVC*, 12:487–498, 1994. 2522, 2526, 2527

[36] Y. Weiss and W. T. Freeman. What makes a good model of natural images? *CVPR*, 2007. 2523, 2524, 2525

[37] A. Yilmaz and M. Shah. Estimation of arbitrary albedo and shape from shading for symmetric objects. *BMVC*, 2002. 2523

[38] R. Zhang, P.-S. Tsai, J. E. Cryer, and M. Shah. Shape-from-shading: a survey. *PAMI*, 21(8):690–706, 2002. 2522, 2527