Color Constancy, Intrinsic Images, and Shape Estimation (Supplementary Material)

Jonathan T. Barron and Jitendra Malik
{barron, malik}@eecs.berkeley.edu

UC Berkeley

1 Priors on Shape

Our prior on shape $f(Z)$ is very similar to that of [1]. It is a linear combination of three costs:

$$f(Z) = \lambda_f f_f(Z) + \lambda_c f_c(Z) + \lambda_k f_k(Z) \tag{1}$$

where $f_f(Z)$ is a "flatness" term that encourages shapes to be fronto-parallel, f_c encourages shapes to face outward at the bounding contour, and f_k is a smoothness term that penalizes variation in mean curvature. The λ multipliers are learned through cross-validation on the training set. Our $f_f(Z)$ and $f_c(Z)$ terms are identical to [1], but we present a refinement to $f_k(Z)$: we make the model single-scale (as multiscale priors don't improve performance significantly with our new multiscale optimization technique), and instead of placing a penalty on the gradient norm of mean curvature, we place the penalty on the pairwise variation of mean curvature between all pixels in a small window, which improves performance. Formally, our prior is:

$$f_k(Z) = \sum_i \sum_{j \in N(i)} \log \left(\sum_{k=1}^{K} \boldsymbol{\alpha}_k \mathcal{N}\left(H(Z)_i - H(Z)_j \,; 0, \boldsymbol{\sigma}_k\right) \right) \tag{2}$$

where $N(i)$ is the 5×5 neighborhood around pixel i, $H(Z)$ is the mean curvature of shape Z, $H(Z)_i - H(Z)_j$ is the difference between the mean curvature at pixel i and pixel j, $K = 40$ (the GSM has 40 discrete Gaussians), $\boldsymbol{\alpha}$ are mixing coefficients, $\boldsymbol{\sigma}$ are the standard deviations of the Gaussians in the mixture. The mean is set to 0, as the most likely shapes should have constant mean curvature (like planes, spheres, cylinders, soap bubbles, etc). The GSM is learned using EM on the training set.

To review, mean curvature is defined as the average of principle curvatures: $H = \frac{1}{2}(\kappa_1 + \kappa_2)$. It can be approximated using the first and second partial derivatives of a surface:

$$H(Z) = \frac{\left(1 + Z_x^2\right) Z_{yy} - 2 Z_x Z_y Z_{xy} + \left(1 + Z_y^2\right) Z_{xx}}{2\left(1 + Z_x^2 + Z_y^2\right)^{3/2}}$$

See [1] for a thorough explanation of how this can be calculated and differentiated efficiently with filter convolutions.

2 Global Reflectance Entropy

Rényi measures of entropy are quadratically expensive to compute naively, so others have used the Fast Gauss Transform [2] and histogram-based techniques [1] to approximate them in linear time. We will generalize the algorithm of [1] for computing 1D entropy to our 3D case.

Given $3 \times n$ matrix $\mathrm{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n]$ we construct a 3D histogram Γ as follows. Γ is initialized to 0, and then for each vector $\mathbf{x}$ we increment the eight surrounding bins in Γ by fractions that sum to 1 using trilinear interpolation:

$$\begin{aligned}
\Gamma\left(\left\lfloor\frac{x_1}{s}\right\rfloor, \left\lfloor\frac{x_2}{s}\right\rfloor, \left\lfloor\frac{x_3}{s}\right\rfloor\right) &+= \left(\left\lceil\frac{x_1}{s}\right\rceil - \frac{x_1}{s}\right)\left(\left\lceil\frac{x_2}{s}\right\rceil - \frac{x_2}{s}\right)\left(\left\lceil\frac{x_3}{s}\right\rceil - \frac{x_3}{s}\right) \\
\Gamma\left(\left\lfloor\frac{x_1}{s}\right\rfloor, \left\lfloor\frac{x_2}{s}\right\rfloor, \left\lceil\frac{x_3}{s}\right\rceil\right) &+= \left(\left\lceil\frac{x_1}{s}\right\rceil - \frac{x_1}{s}\right)\left(\left\lceil\frac{x_2}{s}\right\rceil - \frac{x_2}{s}\right)\left(\frac{x_3}{s} - \left\lfloor\frac{x_3}{s}\right\rfloor\right) \\
&\cdots \\
\Gamma\left(\left\lceil\frac{x_1}{s}\right\rceil, \left\lceil\frac{x_2}{s}\right\rceil, \left\lfloor\frac{x_3}{s}\right\rfloor\right) &+= \left(\frac{x_1}{s} - \left\lfloor\frac{x_1}{s}\right\rfloor\right)\left(\frac{x_2}{s} - \left\lfloor\frac{x_2}{s}\right\rfloor\right)\left(\left\lceil\frac{x_3}{s}\right\rceil - \frac{x_3}{s}\right) \\
\Gamma\left(\left\lceil\frac{x_1}{s}\right\rceil, \left\lceil\frac{x_2}{s}\right\rceil, \left\lceil\frac{x_3}{s}\right\rceil\right) &+= \left(\frac{x_1}{s} - \left\lfloor\frac{x_1}{s}\right\rfloor\right)\left(\frac{x_2}{s} - \left\lfloor\frac{x_2}{s}\right\rfloor\right)\left(\frac{x_3}{s} - \left\lfloor\frac{x_3}{s}\right\rfloor\right)
\end{aligned}$$

Where s is a scale parameter that controls the bin-width of the histogram, and implicitly, accuracy. Using interpolation makes our binning operation smooth and differentiable, which is necessary for calculating the analytic gradient of H with respect to X. With this histogram, we create a new blurred histogram $\tilde{\Gamma} = G * \Gamma$, where G is a three-dimensional Gaussian kernel. This kernel is separable, so we can instead convolve Γ with the a one-dimensional Gaussian kernel $\mathbf{g} = \mathcal{N}(0, \frac{\sqrt{2}\sigma}{s})$ in the three dimensions:

$$\tilde{\Gamma} = \mathbf{g}_1 * (\mathbf{g}_2 * (\mathbf{g}_3 * \Gamma)) \tag{3}$$

With this, the following is a good approximation to our entropy measure:

$$H(\mathrm{X}, \sigma) \approx -\log\left(\sum_{i,j,k} \Gamma_{i,j,k} \times \tilde{\Gamma}_{i,j,k}\right) \tag{4}$$

Which is just the negative log of the inner product between the two histograms. We omit the proof of correctness for brevity's sake, but it looks very similar to the proof of the 1D case in [1]. The analytical derivatives of H with respect to X can be calculated efficiently in the manner of [1].

Note that this formulation is extremely similar to the bilateral grid [3], which is a tool for high-dimensional Gaussian filtering (used mostly for bilateral filtering, hence the name). The calculation of our entropy measure is extremely similar to the "splat, blur, slice" pipeline in other high-dimensional Gaussian filtering works [4], except that after the "slice" operation we take the inner product of the input "signal" and the blurred output signal. This means that we need not actually compute the slice operation, but can instead just compute the inner product directly in the histogram space. This connection means that the body

of work for efficiently computing this quantity in the context of image filtering can be directly adapted to the problem of computing high-dimensional entropy measures. Recent work [4] suggests that for dimensionalities of 3, our bilateral grid formulation is the most efficient among existing techniques, but that this entropy measure could be computed reasonably efficiently in significantly higher-dimensionality spaces (up to 8 or 16) using more sophisticated techniques.

3 Error Metrics

Choosing good error metrics is challenging. Our error metrics are a revision of those in our previous work [1]. We will use the geometric mean of five error metrics: one for shape, one for illumination, one for shading, one for reflectance, and the MIT intrinsic images error metric introduced in [5], which we will refer to as *rs*-MSE (though which the original authors call "LMSE"). We use the geometric mean of these metrics as it is insensitive to the different dynamic ranges of the constituent error metrics, and is difficult to trivially minimize in practice.

Our shape error metric is:

$$N\text{-MSE}(\hat{N}, N^*) = \frac{1}{n} \sum_{x,y} \arccos\left(\hat{N}_{x,y} \cdot N^*_{x,y}\right)^2 \tag{5}$$

This is the mean square error between the angle the normal field $\hat{N}$ of our estimated shape $\hat{Z}$ and the normal field N^* of the ground-truth shape Z^*, in radians. This error metric is invariant to shifts in $\hat{Z}$, but is sensitive to other errors.

For illumination, our error metric is:

$$L\text{-MSE}(\hat{L}, L^*) = \frac{1}{n} \min_{\alpha} \sum_{x,y} ||\alpha V(\hat{L})_{x,y} - V(L^*)_{x,y}||_2^2 \tag{6}$$

Which is the scale-invariant MSE of a rendering of our recovered illumination $\hat{L}$ and the ground-truth illumination L^*. $V(L)$ is a function that renders the spherical harmonic illumination L on a sphere and returns the log-shading. $V(L)_{x,y}$ is a 3-vector of log-RGB at position (x, y) in the renderings. The α multiplier makes this error metric invariant to absolute scaling, meaning that estimating illumination to be twice as bright or half as bright doesn't change the error. But because there is only one multiplier rather than individual scalings for each RGB channel, this error metric is sensitive to the overall color of the illuminant. This choice seems consistent with what we would like: estimating absolute intensity of an illuminant from a single image is both incredibly difficult and not very useful, but estimating the color of the illuminant is a reasonable thing to expect from an algorithm, and would be useful for many applications (color constancy, relighting, reflectance estimation, etc). We impose our error metric in the space of visualizations of the illumination rather than in the space of the actual

spherical harmonic coefficients that generated that visualization, both because it makes our error metric invariant to the choice of illumination model, and because we found that often the recovered illumination could look quite similar to the ground-truth, while having a very different spherical harmonic representation.

For shading and reflectance, we use:

$$s\text{-MSE}(\hat{s}, s^*) = \frac{1}{n}\min_{\alpha} \sum_{x,y} \left\|\alpha\hat{s}_{x,y} - s^*_{x,y}\right\|_2^2 \tag{7}$$

$$r\text{-MSE}(\hat{r}, r^*) = \frac{1}{n}\min_{\alpha} \sum_{x,y} \left\|\alpha\hat{r}_{x,y} - r^*_{x,y}\right\|_2^2 \tag{8}$$

These are the scale-invariant MSEs of our recovered shading $\hat{s} = \exp(S(\hat{Z}, \hat{L}))$ and reflectance $\hat{r} = \exp(\hat{R})$. Just like in L-MSE, we are invariant to absolute scaling of all RGB channels at once, but not invariant to scaling each channel individually. This makes these error metrics sensitive to errors in estimating the overall color of the shading and reflectance images, but invariant to illumination. Note that these error metrics are of shading and reflectance, not of log-shading and log-reflectance, even though the rest of this paper is written almost entirely in terms of log-intensity. We could have used shift-invariant error metrics in log-intensity space, but we found these to be too sensitive to errors in dark regions of the image — places in which we'd expect any algorithm to do worse, simply because there is less signal.

Our final error metric is the metric introduced in conjunction with the MIT intrinsic images dataset [5], which the authors refer to as LMSE, but which we will call rs-MSE to minimize confusion with L-MSE. This metric measures error for both reflectance and shading, and is locally scale-invariant. The intent of the local scale-invariance is to make the metric insensitive to low-frequency errors in either shading or reflectance. In keeping with this spirit, we apply this error metric individually to each RGB channel and take the mean of those three errors as rs-MSE, making this error metric not just robust to low-frequency error, but robust to most errors in estimating the color of the illumination. This error metric therefore serves to be somewhat complementary to s-MSE and r-MSE, which are sensitive to everything except absolute intensity.

References

1. Barron, J.T., Malik, J.: Shape, albedo, and illumination from a single image of an unknown object. CVPR (2012)
2. Finlayson, G.D., Drew, M.S., Lu, C.: Entropy minimization for shadow removal. IJCV (2009)
3. Chen, J., Paris, S., Durand, F.: Real-time edge-aware image processing with the bilateral grid. SIGGRAPH (2007)
4. Adams, A., Baek, J., Davis, M.A.: Fast high-dimensional filtering using the permutohedral lattice. Eurographics (2012)
5. Grosse, R., Johnson, M.K., Adelson, E.H., Freeman, W.T.: Ground-truth dataset and baseline evaluations for intrinsic image algorithms. ICCV (2009)

6. (http://www.hdrlabs.com/sibl/archive.html)
7. Gehler, P., Rother, C., Kiefel, M., Zhang, L., Schoelkopf, B.: Recovering intrinsic images with a global sparsity prior on reflectance. NIPS (2011)
8. Shen, J., Yang, X., Jia, Y., Li, X.: Intrinsic images using optimization. CVPR (2011)
9. Tappen, M.F., Freeman, W.T., Adelson, E.H.: Recovering intrinsic images from a single image. TPAMI (2005)
10. Horn, B.K.P.: Determining lightness from an image. Computer Graphics and Image Processing (1974)
11. Gijsenij, A., Gevers, T., van de Weijer, J.: Generalized gamut mapping using image derivative structures for color constancy. IJCV (2010)

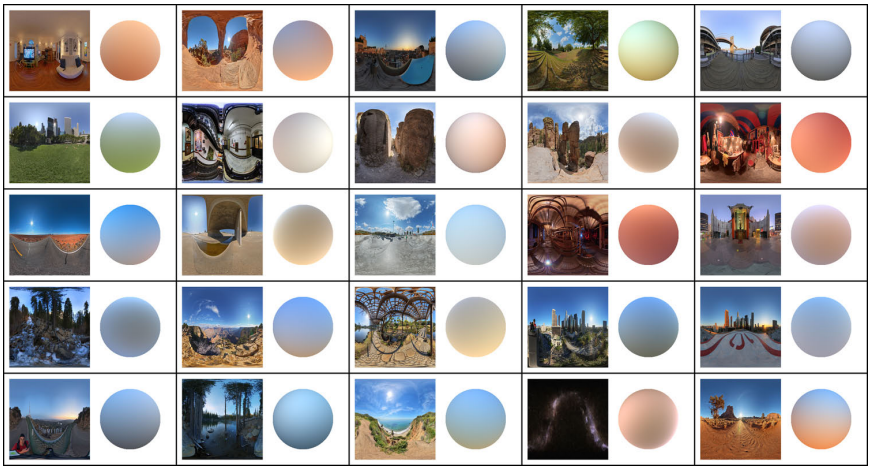

Fig. 1. Some examples of the environments from the sIBL Archive [6], rendered in lat-lon coordinates, with the corresponding spherical harmonic illuminations which we recover from the environment JPEGs. We see that natural illumination displays much more color than is present in the MIT dataset.

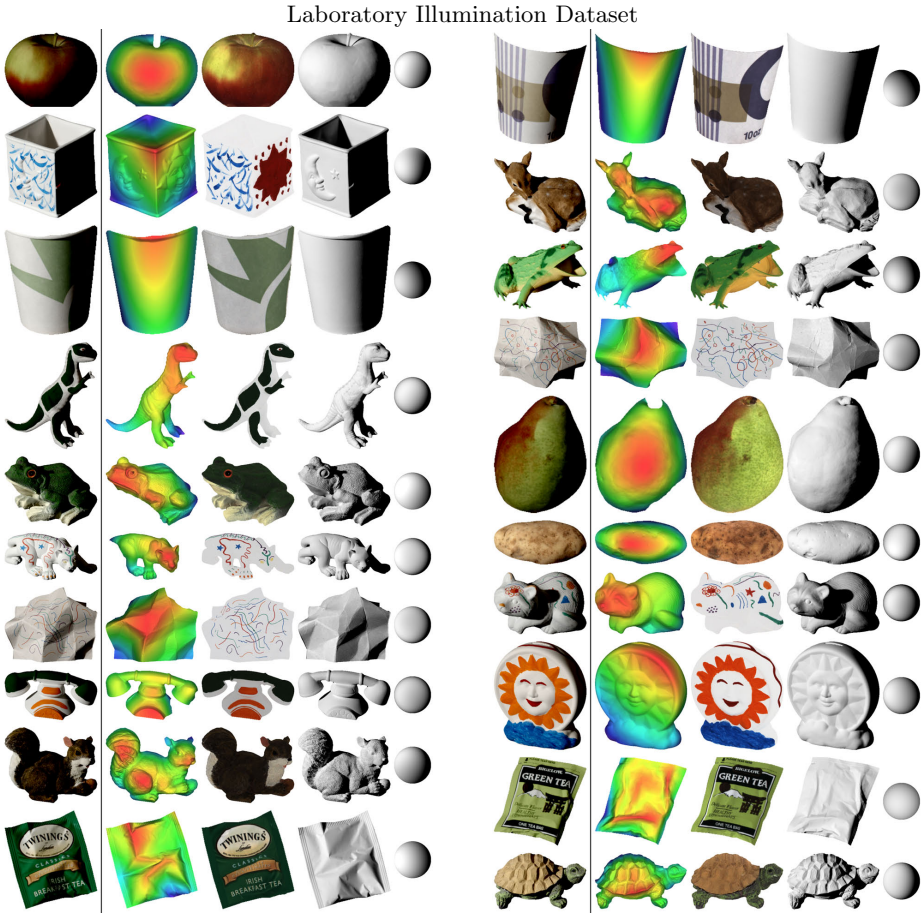

Fig. 2. The twenty objects in the MIT Intrinsic Images dataset [5], with the ground-truth shape and illumination provided by [1]. Our training set is to the left, and the test set is to the right. We see that the shading images are monochrome, and that the illuminations are nearly entirely white. It is because of the white, "laboratory"-like nature of these illuminations that we introduce our novel dataset, shown in Figure 3.

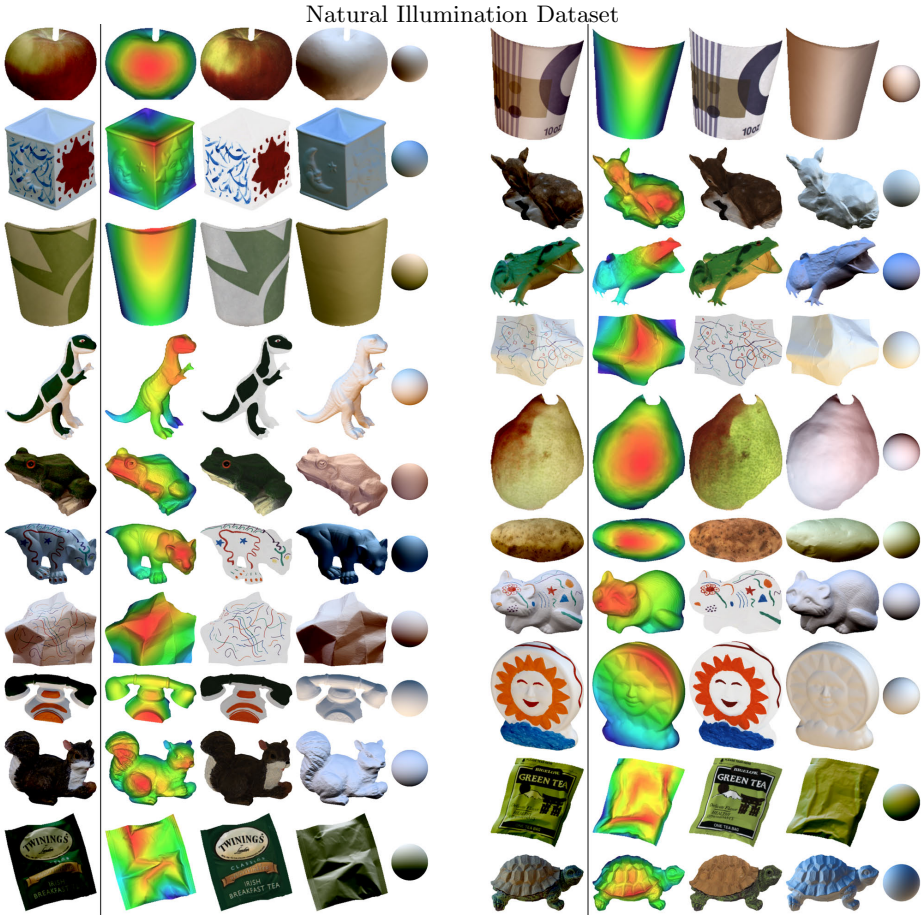

Fig. 3. Here we have taken the ground-truth shape and reflectances from the MIT Intrinsic Images dataset [5, 1], and rerendered them under illuminations produced from environment maps taken from the sIBL Archive [6]. The resulting shading images look much more natural and colorful than in the "laboratory" dataset in Figure 2. Some of the images are missing, because the corresponding ground-truth illumination is missing wherever the photometric stereo algorithm of [1] failed when the dataset was being created, mostly due to cast and attached shadows. Images are rendered with increased contrast for improved visibility, but the contrast is equivalent in this visualization and in Figure 2, these illuminations are just must softer than those of the original MIT dataset.

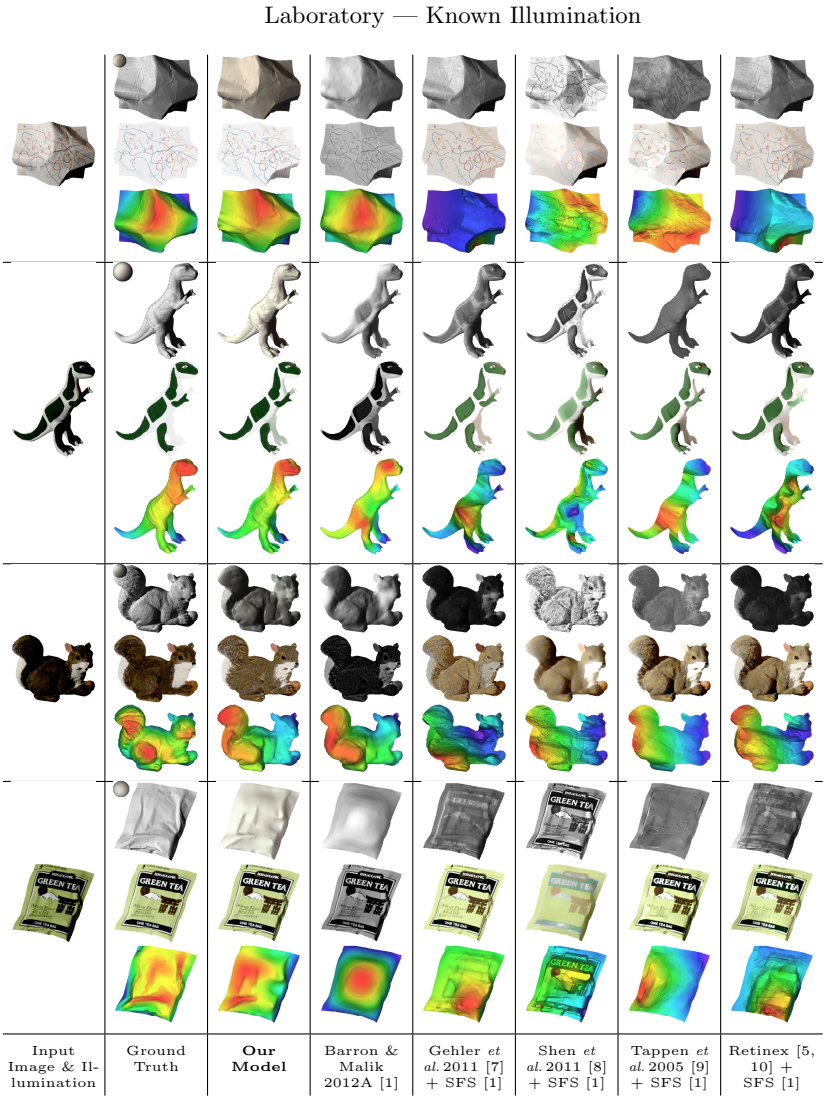

Fig. 4. The output of our algorithm, and others, for the task of recovering shape and reflectance given a single image, and known illumination, using the "Laboratory" illumination version of the MIT intrinsic images dataset. Our model outperforms all other algorithms. Comparing our model to the grayscale model of Barron and Malik shows that using color improves results. The top-performing intrinsic image algorithm of Gehler *et al.* often produces excellent looking shading images, but when these shading images are fed to a shape-from-shading algorithm, the resulting shape tends to look very bad, probably because the shading image is inconsistent with the known illumination.

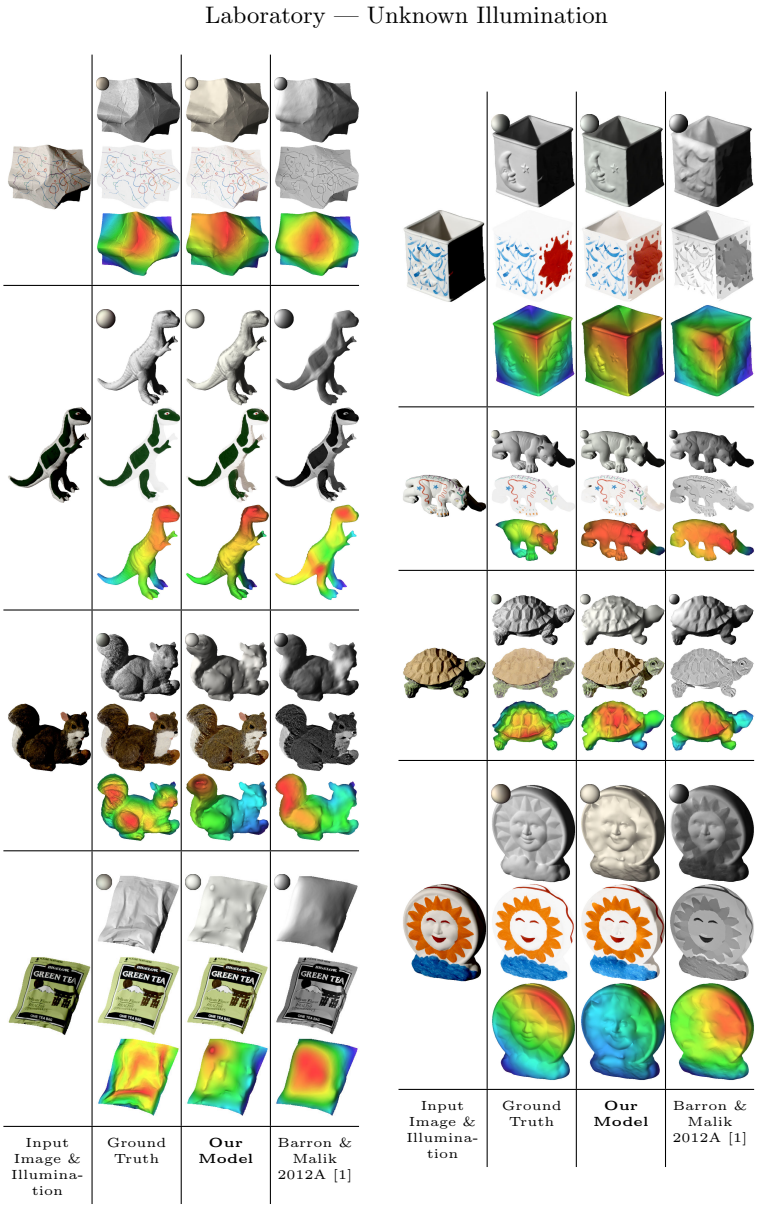

Fig. 5. The output of our algorithm, and others, for the task of recovering shape and reflectance given a single image, and unknown illumination, using the "Laboratory" illumination version of the MIT intrinsic images dataset. . Color appears to be more helpful for the man-made objects than for more "natural" objects (such as the squirrel or turtle).

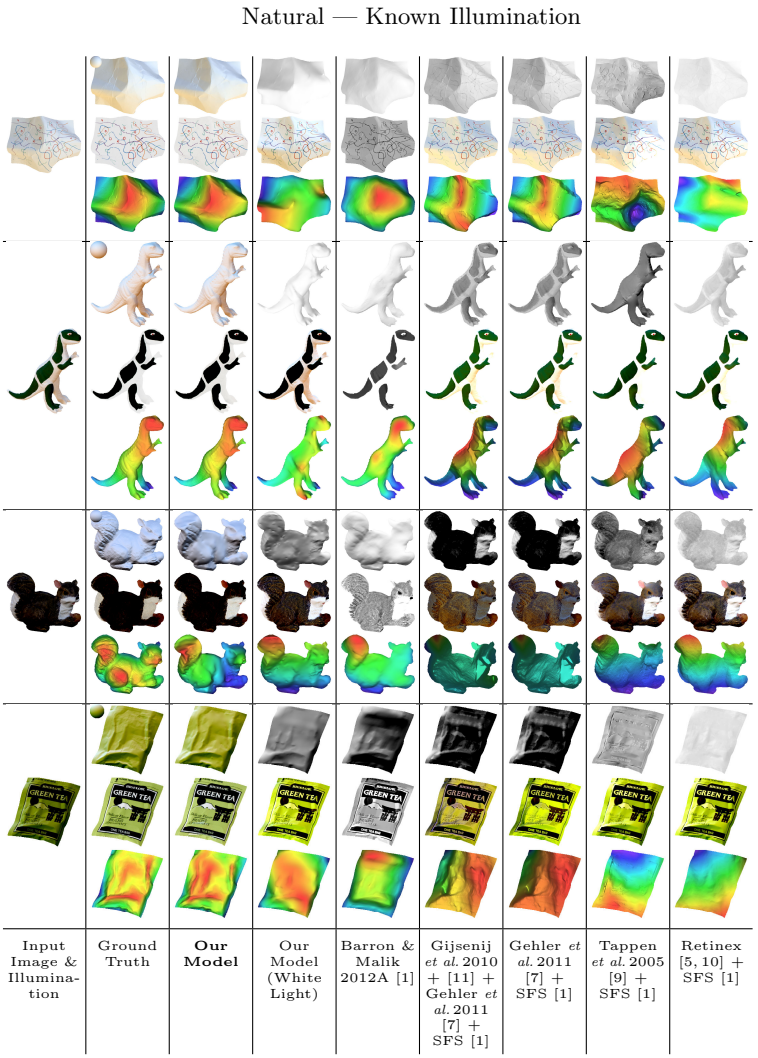

Fig. 6. The output of our algorithm, and others, for the task of recovering shape and reflectance given a single image, and known illumination, using our "Natural" illumination version of the MIT intrinsic images dataset. Our results are very good, and often indistinguishable from ground-truth. If illumination is assumed to be white, performance worsens tremendously. The grayscale technique of Barron and Malik performs reasonably, as it is not confused by color, but produces significantly less precise shapes than our model. The technique of Gehler *et al.*, which worked well in the laboratory illumination case, performs very poorly here, as its assumption of a white light is violated, and the color illumination therefore confuses it. This is an issue whether or not a contemporary white balance algorithm [11] is run on the input image beforehand. The other intrinsic image algorithms are similarly confused by natural illumination.

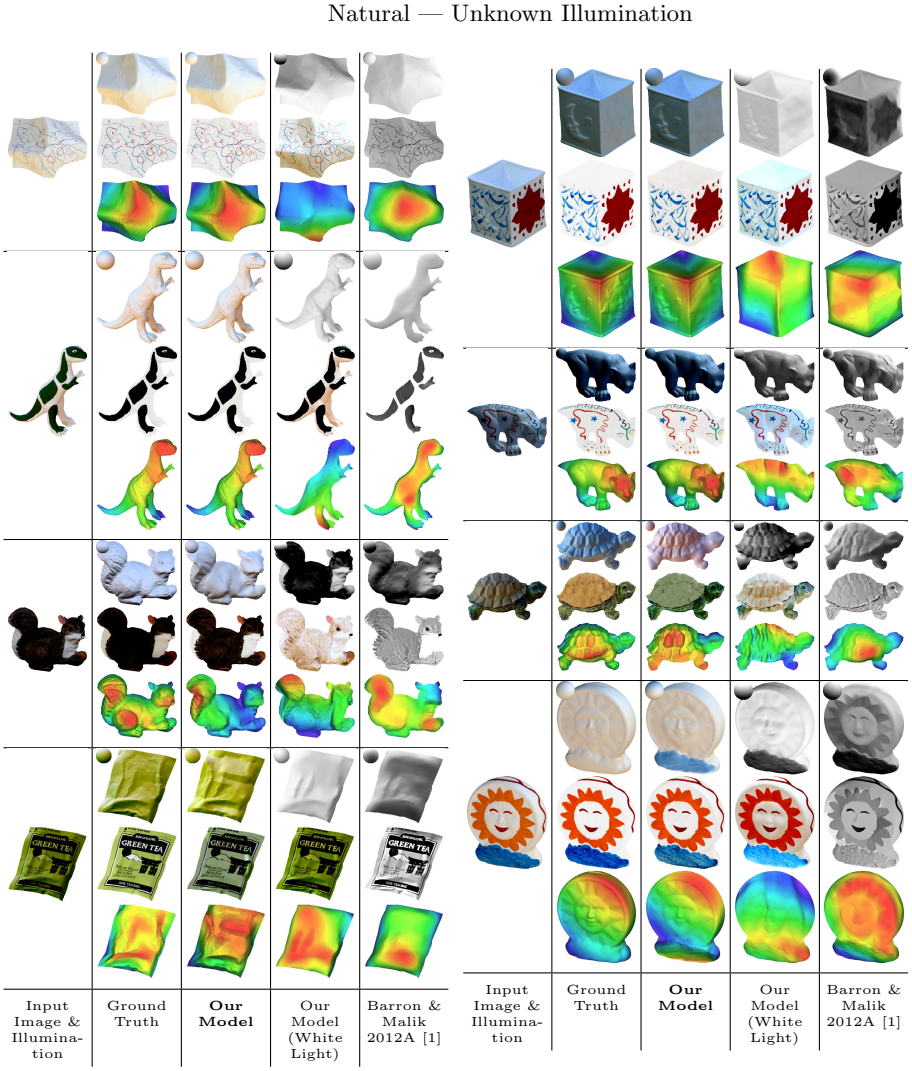

Fig. 7. The output of our algorithm, and others, for the task of recovering shape and reflectance given a single image, and unknown illumination, using our "Natural" illumination version of the MIT intrinsic images dataset. Results are often good whether or not illumination is known, but errors are more likely in the unknown illumination case, as we would expect. If we either assume white illumination performance drops substantially, to such a degree that the grayscale method of Barron and Malik 2012A produces better results, as it is not mislead by this color information.

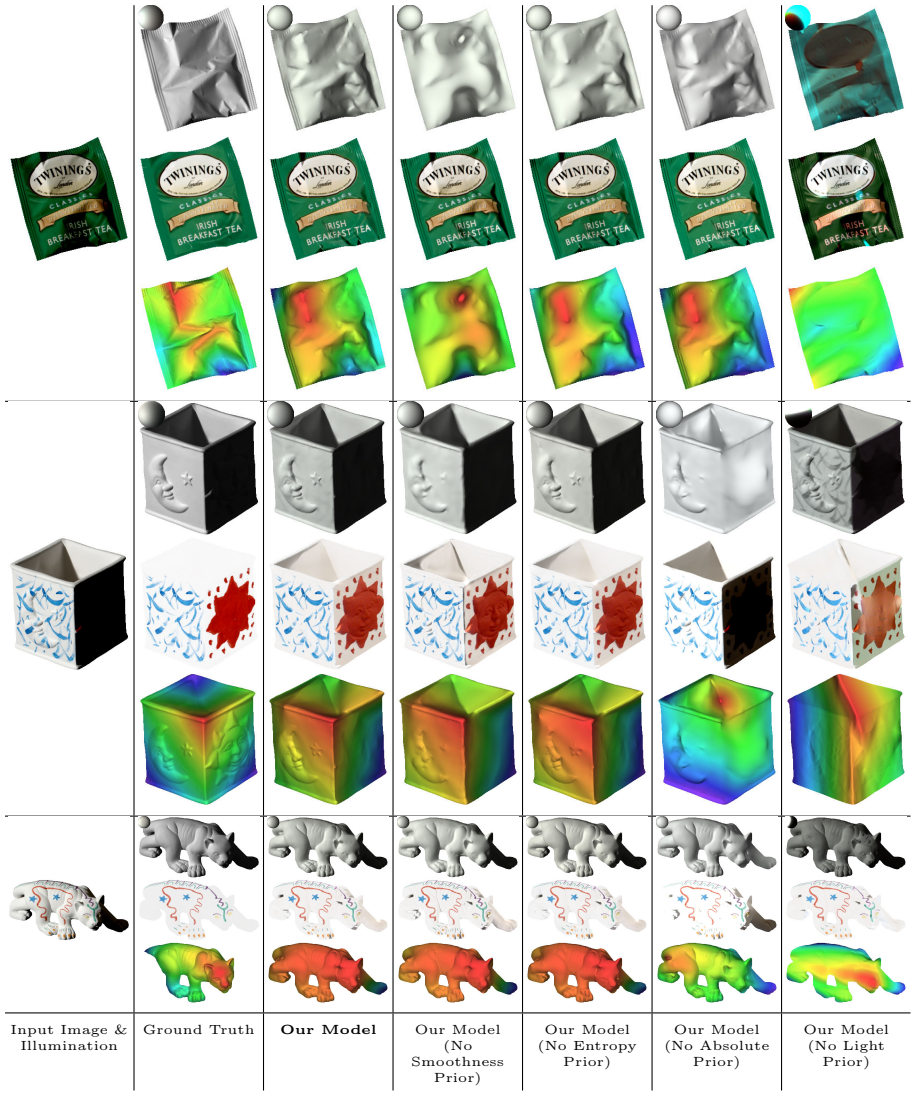

Fig. 8. Our ablation study. Omitting the smoothness prior gives results that are reasonable at a coarse scale, but missing fine-scale variations in shape (which are local phenomena). Omitting the entropy prior does a good job of assigning edges to either shape or reflectance, but produces a reflectance image which slowly varies across the entire object, as opposed to being composed of a few globally consistent colors. Omitting the absolute prior results in strange-looking or impossible colors in the reflectance image. Omitting the prior on illumination results in absurdly flattened shapes and harsh illuminations — a very extreme member of the bas-relief family.