Discriminative Shape from Shading in Uncalibrated Illumination

Stephan R. Richter   Stefan Roth
Department of Computer Science, TU Darmstadt

Abstract

Estimating surface normals from just a single image is challenging. To simplify the problem, previous work focused on special cases, including directional lighting, known reflectance maps, etc.*, making shape from shading impractical outside the lab. To cope with more realistic settings, shading cues need to be combined and generalized to natural illumination. This significantly increases the complexity of the approach, as well as the number of parameters that require tuning. Enabled by a new large-scale dataset for training and analysis, we address this with a discriminative learning approach to shape from shading, which uses regression forests for efficient pixel-independent prediction and fast learning. Von Mises-Fisher distributions in the leaves of each tree enable the estimation of surface normals. To account for their expected spatial regularity, we introduce spatial features, including texton and silhouette features. The proposed silhouette features are computed from the occluding contours of the surface and provide scale-invariant context. Aside from computational efficiency, they enable good generalization to unseen data and importantly allow for a robust estimation of the reflectance map, extending our approach to the uncalibrated setting. Experiments show that our discriminative approach outperforms state-of-the-art methods on synthetic and real-world datasets.*

1. Introduction

Estimating surface normals from just a single image is a severely under-constrained problem. Previous work has thus made a number of simplifying assumptions, *e.g.*, presuming smooth surfaces, uniform albedo, and directional lighting from a single light source of known direction. Yet, relying on a single directional light, or assuming a given reflectance map renders shape from shading impractical. To go beyond the lab, some of these assumptions need to be relaxed. We thus investigate estimating the reflectance map of a diffuse object with uniform albedo together with its surface under uncontrolled illumination, given only a single image (Fig. 1). Moreover, we aim to avoid strong spatial regularization to recover fine surface detail. To address this challenging setting, shading cues need to be generalized to more realistic lighting and also combined due to their complementary strengths. This increases the model and computational complexity, as well as the number of parameters.

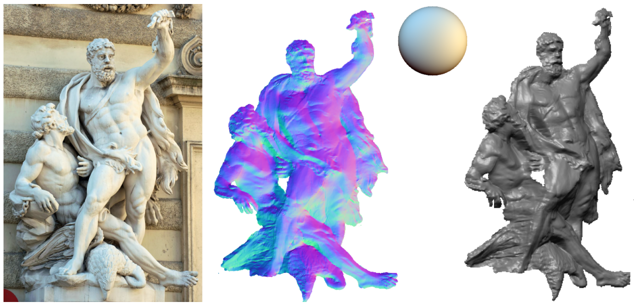

Figure 1. Qualitative results for shape and reflectance estimation from a single image: input image [30], estimated normals and reflectance map from our method, and novel view (from left to right).

We approach these challenges with a discriminative learning approach to shape from shading. This notably makes it easy to combine several shading cues. We begin by considering *(1) color*, which helps under hued illumination [18], and can be exploited with a second order approximation to Lambertian shading [25]. While powerful, our experimental investigation (Sec. 8.1) shows that correlated color channels (*e.g.*, in near white light) or the presence of noise severely impair accuracy. We thus add *(2) local context*, which aids disambiguation [33], but until now has been limited to directional lighting. Instead of using colors from a neighborhood directly, we choose a Texton filter bank [28] for capturing local context. Employing these filters in a learning framework allows for automatic adaptation to uncontrolled lighting and fine surface detail. We finally exploit *(3) silhouette features*. Previous work has constrained surface normals at the occluding contour [16, 20], and propagated this information to the surface interior during global reasoning. We generalize the occluding contour constraint to the surface interior and provide (spatial) contour information at every pixel. These silhouette features yield a coarse estimate of the surface from just the silhouette, which we additionally exploit for estimating the reflectance map.

Adopting a learning approach to uncalibrated shape from shading poses several challenges: First, example-based

To appear in the Proc. of the *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, June 2015.

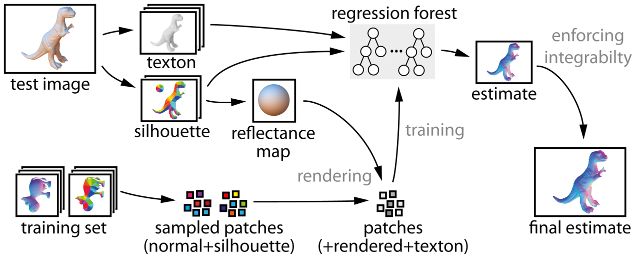

Figure 2. Pipeline for training and testing. For each test image, we estimate a reflectance map to train the regression forests on synthetically generated data. We optionally enforce integrability on their pixel-wise predictions.

approaches require a database of surfaces imaged under the same conditions as the object in question. Capturing all possible combinations of surfaces and lighting conditions is next to impossible, and placing known objects in the scene [14] often impractical. Recent learning approaches [5, 19] thus created a database on-the-fly by rendering synthetic shapes under a known lighting condition. We follow this avenue, but capture the variation of realistic surfaces with a significantly larger database of 3D models created by artists. Second and unlike [5, 19], we aim to cope with unknown illumination at test time. We address this by estimating the reflectance map from silhouette features, and only then training the discriminative approach on the estimated reflectance. Third, (re-)training a model for a lighting condition at test time requires efficient learning and inference. Adapting regression forests to store von Mises-Fisher distributions in the leaves, as well as leveraging the set of cues discussed above enables us to perform efficient pixel-independent surface normal prediction. To further refine the estimate, we optionally enforce integrability on the predicted surface. Fig. 2 depicts the entire pipeline.

To assess the contribution of the different cues, we first evaluate their importance statistically and as a component of our approach. Moreover, we evaluate and compare our method on both synthetic and real-world data, where it outperforms several state-of-the art algorithms.

2. Related Work

Shape from shading has an extensive literature [9, 34], limiting us to the most relevant, recent work here. For decades, Lambertian shape from shading has been studied under the assumption of a single white point light source, presuming that this simplifies the problem. However, chromatic illumination not only resembles real-world environments much better, it also constrains shape estimation significantly, permitting substantially increased performance [15, 18]. Nevertheless, these methods rely heavily on favorable illumination and neglect to exploit shape statistics under nearly monochromatic lighting. Their assumption of known illumination further limits practical applicability.

In addition to eliminating the point light assumption, recent work aimed to infer the shape jointly with material properties or illumination. Oxholm and Nishino [22] estimated shape together with the object's bidirectional reflectance function (BRDF) by exploiting orientation cues present in the lighting environment. However, a high-quality environment map needs to be captured. Barron and Malik [1, 2] estimate shape as part of a decomposition of a single image into its intrinsic components. Since they optimize a generative model, training and inference take significant time and extending the model with additional cues is complicated. Moreover, strong regularity assumptions need to be made, limiting the recovery of fine surface detail.

Learning approaches to shape from shading have so far been limited by the lack of suitable training data. They have relied on range images or synthetic data [3, 19, 31], both causing problems of their own: While the noise of range images is a severe problem in predicting fine-grained surface variations, synthetic datasets often fail to capture the variation of real-world environments. Recently, Khan *et al*. [19] trained a Gaussian mixture model on the isophotes, using synthetic data and a database of laser scans. Barron and Malik [1] learned their shape priors from half the MIT intrinsic image dataset [13]. In addition, example-based methods [5, 23] have shown reasonable qualitative performance; quantitative results have not been reported though. While learning methods tend to perform better than their hand-tuned counterparts, they have been limited by simplistic shape priors and the lack of adequate training data.

The work of Hertzmann and Seitz [14] also bears some relations, which used objects of known geometry imaged under the same illumination to reconstruct shape from photometric stereo. Multiple images are required and a known object needs to be present in each image. In contrast, we only need one image of an unknown object and use "example geometry" only to synthesize our training data.

Our regression tree-based predictor is most related to Geodesic Forests [21]; both approaches make pixel-independent predictions by incorporating spatial information directly into the tree-based approach. However, Kontschieder *et al*. [21] use it for discrete labeling tasks, such as semantic segmentation, have a complex entanglement, and use generalized geodesic distances as spatial features. We instead predict a two-dimensional continuous variable (normal direction), employ the proposed silhouette features, and use only a single stage.

3. Overview

We begin with an overview of our discriminative approach (Fig. 2). Given a test image of a diffuse object with uniform albedo, we first extract color, textons and our silhouette features (Sec. 6). Based on the silhouette features, we estimate the reflectance map (Sec. 7), with which in turn we render patches of objects from our database. The features (same as for testing) and surface normal of the central pixel of each synthetic patch form the training dataset, on which the regression forest (Sec. 5) is trained. After training, the forest predicts surface normals for each pixel of the test image from the features we extracted at the beginning. Optionally, we enforce integrability of the normal field.

4. Data for Analysis and Training

High-quality training data for learning surface variation has been scarce. With the advent of low-cost depth sensors, the situation has improved, but real world range images are often too noisy to learn fine-grained structures. Alternatively, large amounts of synthetic data have been generated, which often resembles simple geometric shapes like cylinders or blobs [3, 5, 18, 31]. However, the employed parametric models often fail to capture real-world surface variations, like self-occlusions or wrinkles from clothing.

By instead leveraging a dataset of shapes created by an artist [8], we combine the advantages of both range maps and synthetic data: First, being manually created by a modeling expert, the shapes resemble real-world objects with complex phenomena like self-occlusions. Second, having a large set of 3D models allows to render them in different orientations and to generate a virtually infinite set of training images. The 3D models we use cover a range of categories, mainly with an organic shape, such as humans and animals.

It is important to note that our dataset is much larger (100 objects) and more varied than the ones considered in other learning approaches to shape from shading. [19] used only 6 realistic surfaces of the same object class (faces), while [2] used 10 objects by taking half of the MIT intrinsic image dataset for training (the other for testing). Although we trained our algorithm on a dataset that is qualitatively rather different from all test datasets, we achieve state-of-the-art performance for a variety of test datasets (see Sec. 8).

5. Discriminative Prediction of Normals

Motivated by the success of decision and regression tree-based methods in diverse areas, such as human pose estimation [27], image restoration [17], or semantic labeling [21], we propose to use regression forests for shape from shading. Note that we only describe the basic learning approach here, and defer a more thorough discussion of the used features to Sec. 6. A key motivation for choosing regression forests is that learning and prediction steps are computationally efficient, a crucial requirement as both learning and prediction have to be carried out at test time once the reflectance map has been estimated (Fig. 2). The different trees of the forest can be trained in parallel. Moreover, as we predict the surface normals of objects independently for each pixel, estimation is efficient and parallelizable. Since the prediction does not necessarily result in an integrable normal field, integrability can be enforced in a post-processing step.

Basic regression forest model. Regression forests form a prediction of the output variable (here, at a single pixel) by traversing a tree, such that the traversal path allows the model to choose an appropriate prediction based on the input features [4]. To that end, each (non-leaf) node has a split criterion, here a threshold on one input feature, as usual. Depending on the feature response, the traversal follows the left or right branch, until a leaf is reached. The leaves each store a probability distribution over the output variable, which ultimately enables prediction. For robustness, the predictions of several trees in a forest are averaged.

Normal vectors as output. Predicting normal vectors comes with a unique set of challenges. First, the output is a continuous variable, which is normally addressed by storing in each leaf the average of all training samples that have been associated with that particular leaf. This can be thought of as storing the mean of a multivariate Gaussian, which is sensible when the posterior is reasonably approximated by a Gaussian distribution in $\mathbb{R}^d$. The second challenge in predicting surface normals is that they are distributed on a 3-dimensional unit hemisphere; thus a Gaussian assumption is not appropriate. We address this by modeling the prediction in each leaf as a von Mises-Fisher distribution [10]; we store the mean and dispersion parameters.

More specifically, the von Mises-Fisher distribution models unit vectors on a d-dimensional hypersphere. In our case of $d = 3$, its density function is given as

$$p(\mathbf{n}; \boldsymbol{\mu}, \kappa) = \frac{\kappa}{2\pi(e^{\kappa} - e^{-\kappa})} \exp(\kappa \boldsymbol{\mu}^{\mathrm{T}} \mathbf{n}), \tag{1}$$

where the normal $\mathbf{n} \in S^2$ is distributed around a mean vector $\boldsymbol{\mu}, ||\boldsymbol{\mu}|| = 1$ with a dispersion $\kappa \in \mathbb{R}$ (analogous to the precision of a Gaussian distribution).

Learning. Learning proceeds mostly as for standard regression forests. For every tree of the forest, we randomly choose a 90% subset of the training data. During learning, for each node a random set of features is chosen from which the best is picked as split criterion. The feature chosen is the one that minimizes the aggregated entropy of the the new child nodes compared to the entropy of their parent node. The entropy is calculated from the distribution of the output variable.

We estimate the parameters of the von Mises-Fisher distributions following Dhillon and Sra [7], who showed that the maximum likelihood estimate is approximated well by

$$\hat{\boldsymbol{\mu}} = \frac{\mathbf{r}}{||\mathbf{r}||} \quad \text{and} \quad \hat{\kappa} = \frac{3\bar{R} - \bar{R}^3}{1 - \bar{R}^2}, \tag{2}$$

where $\mathbf{r} = \sum_{i=1}^{N} \mathbf{n}_i$ is the resultant vector, and $\bar{R} = \frac{||\mathbf{r}||}{N}$ is the average resultant length.

As is common for most learning-based methods in shape from shading [5, 19], each unseen reflectance map requires re-training the model. To do this efficiently, we sample 5×5 normal patches from our dataset (Sec. 4) and store normals and silhouette features ahead of time, as they are independent from the lighting condition. When we encounter a new illumination condition, the normals are rendered, the remaining features computed, and the forests trained. Rendering and feature extraction takes less than 1 second for the whole dataset; training takes about 90 seconds and inference less than a second. Note that an important difference to [5, 19] is that we do not require the lighting at test time to be known, but rather estimate it as well (Sec. 7).

We create 10 training images for each of the 100 models in our dataset by placing an (orthographic) camera such that it points to the model center from random positions. We evaluated several numbers of patches on a validation set (different from the models used for testing) and found that 100–200 samples per image give the best trade-off between performance and time needed for training.

Integrability. The regression forests output a surface normal for each pixel independently of the neighboring predictions. Without any spatial regularization, predicted surfaces are more susceptible to image noise; yet penalizing discontinuities usually results in oversmoothed surfaces. Therefore, when fusing pixel-independent predictions, we only enforce integrability, as it is necessary for deriving a valid surface. Integrability requires the surface normal derivatives to fulfill

$$\frac{\partial^2 z}{\partial u \partial v} = \frac{\partial^2 z}{\partial v \partial u}, \tag{3}$$

where the depth z is a function of the image coordinates u, v. To penalize violations of Eq. (3), several approaches have been proposed, *e.g.* [24, 26]; we evaluate different choices in Sec. 8.2.

6. (Spatial) Features

Shape from shading, like other pixel labeling / prediction problems, benefits from taking into account the spatial regularity of the output, or in other words modeling the expected smoothness of the recovered surface. That is, neighboring predictions should account for the fact that their normals are often very similar. Regression forests [4], which we use here, perform pixelwise independent predictions and thus do not necessarily model such regularities well. To address this, regression tree fields [17] predict the parameters of a Gaussian random field instead of the output variables directly; the prediction is obtained by maximum a-posteriori (MAP) estimation in the specified conditional random field. One drawback is that the MAP estimation step incurs a computational overhead, which also makes learning inefficient. Inspired by geodesic forests [21], which include a geodesic distance feature to sidestep explicit modeling of more global dependencies of the output, we here aim to additionally devise *spatial features* that allow promoting spatial consistency despite pixel-independent prediction.

Basic color feature. Regression trees benefit from input features that strongly correlate with the desired output [4], because a strong correlation allows for splits that reduce the entropy well. Possibly the strongest correlation exists between the surface normal and the color; their relation is described by the rendering equation. In the common Lambertian case, we can take a second order approximation [25]: $I_c = \hat{\mathbf{n}}^{\mathrm{T}} \mathbf{M}_c \hat{\mathbf{n}}$ at each point on the surface, where I_c is the intensity of color channel c, $\hat{\mathbf{n}}$ is the surface normal in homogeneous coordinates, and $\mathbf{M}_c$ is a symmetric 4×4 matrix representing the reflectance map for that color channel. A single input image thus puts 3 nonlinear constraints onto the 2 unknowns of the surface normal at each pixel. Under ideal circumstances, the reflectance maps are independent from each other and thus produce small isophotes (areas with the same luminance), such that a surface can be recovered very well with just the color. In this case shape from shading becomes similar to photometric stereo [18]. However, if the reflectance maps and corresponding constraints are more correlated, *e.g.* in nearly white light, large isophotes result in many surface patches that explain the same color. Moreover, the problem is exacerbated by image noise. Hence, to avoid making strong assumptions about the type of lighting present, we not only consider the color, but look for spatial features that depend on a neighborhood of pixels as well as the object contour, and are able to reduce the remaining ambiguity, even in the absence of an explicit spatial model.

Texton features. To capture how the local variation of the input image correlates with the output, we first compute features from a texton filter bank [28]. Texton filters consist of Gaussians, their derivatives, as well as Laplacians at multiple scales, and have been used in many areas, such as material classification, segmentation, and recognition. While having been used in shape from texture [32], to our knowledge this is the first application to shape from shading. Before filtering we convert the image to the L*a*b* opponent color space. Gaussian filters are computed on all channels, while the remaining filters are applied only to the luminance channel. As we will see below, the local context from texton features leads to a strong increase in accuracy compared to using color alone, as their embedding in a discriminative learning framework allows for adaption to various types of surface discontinuities instead of simply assuming smoothness as has been common in shape from shading.

Magnifying the local context by enlarging the filters can lead to faster convergence to an integrable surface, but requires a larger dataset to capture fine detail and achieve similar generalization. In our experiments, we used filters that match the normal patches in size (5×5).

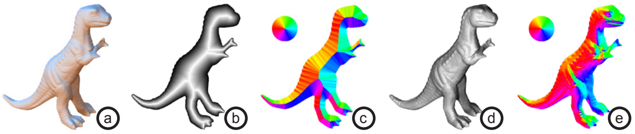

Figure 3. Objects that are convex or composed of convex parts *(a)* exhibit a strong correlation between silhouette-based features like relative distance *(b)* or direction to the silhouette *(c)* and out-of-plane *(d)* and in-plane components *(e)* of their surface normals.

6.1. Silhouette features

Projected onto the image plane, normals are not distributed equally across the object. Most objects are roughly convex, or composed of convex parts. Thus, normals at the center of an object tend to face the viewer and normals at the occlusion boundary face away from the viewer [16, 20]. Consequently, the probability of a normal facing a certain direction is not uniform given a position within the projection. Previous work has exploited this fact only by placing priors on the normals at the occlusion boundary and propagating information to the interior with a smoothness prior [22]. As both priors do not consider scale, balancing them can be challenging, even within the same object, as it may contain parts of different scale (*e.g.*, the tail *vs.* head of the dinosaur in Fig. 3). Here, we consider a relation between the silhouette and the normal that is more explicit and automatically adapts to scale.

To that end, let us first look at the correlation between a point's surface orientation and its position within the object's projection onto the image plane. Consider the object in Fig. 3. As expected [16, 20] and can be seen in the visualization of the out-of-plane component *(d)* (white – toward the viewer, black – away), normals are orthogonal to the viewing direction starting at the silhouette. Moving inwards the normals change until they finally face the viewer. If we now look further at the distance of an interior point to the silhouette *(b)*, we can see some apparent correlation. Similarly, we can see apparent correlation between the direction to the nearest point on the silhouette *(c)* and the image-plane component of the normal *(e)*. We now formalize and analyze this relationship.

If B denotes the set of points on the occlusion boundary, we define the absolute distance of an interior point $\mathbf{p}$ to the contour as

$$d_{\text{abs}}(\mathbf{p}) = \min_{\mathbf{b}\in B} ||\mathbf{p}-\mathbf{b}||. \tag{4}$$

However, we are not interested in the absolute distance, as it depends on the scale of the object. To make it scale-invariant, we normalize it by the length of the shortest line segment that passes through $\mathbf{p}$ and connects boundary and the medial axis of the object. The medial axis is the set of all points that have 2 closest points on the boundary. If M denotes the medial axis and $\overline{\mathbf{pb}}$ the (infinite) line that passes through $\mathbf{p}$ and $\mathbf{b}$, we define the relative distance to the silhouette as

$$d_{\text{rel}}(\mathbf{p}) = \min_{\mathbf{b}\in B} \min_{\mathbf{m}\in M\cap\overline{\mathbf{pb}}} \frac{||\mathbf{p}-\mathbf{b}||}{||\mathbf{m}-\mathbf{b}||}, \tag{5}$$

i.e. the relative distance is normalized by the minimal line that passes through $\mathbf{p}$ and connects medial axis and contour. In practice, we approximate Eq. (5) using two distance transforms, d_B for the contour set and d_M for the medial axis. We thus define the scale-invariant boundary distance

$$d'_{\text{rel}}(\mathbf{p}) = \frac{d_B(\mathbf{p})}{d_B(\mathbf{p}) + d_M(\mathbf{p})}. \tag{6}$$

Finally, we define the direction to the contour as

$$\beta(\mathbf{p}) = -\frac{\nabla d'_{\text{rel}}(\mathbf{p})}{||\nabla d'_{\text{rel}}(\mathbf{p})||}. \tag{7}$$

Statistical analysis. To analyze the correlation between the position relative to the silhouette and surface orientation, we calculated the relative distance and direction to the silhouette for multiple datasets and plotted them against the out-of-plane and the image-plane component of the surface normals (Fig. 4). We analyze three different datasets: synthetic data ("blobby shapes") [18] in the first column, real world data from the MIT intrinsic image dataset [13] in the second column, and a collection of 3D models generated by artists in the third column. Across all datasets we observe a strong correlation between the plotted variables. The direction to the silhouette has a strong linear relation to the image-plane component; the relative distance has a quadratic relation to the out-of-plane component. This clearly suggests that the proposed silhouette features should be useful for surface reconstruction from a single image. We evaluate the importance of our input features for surface prediction in Sec. 8.1.

7. Reflectance Map Estimation

Assuming distant light sources and no self-reflections or occlusions, all observable reflectance values of an object with uniform albedo can be mapped one-to-one onto a hemisphere. Moreover, a Lambertian reflectance map can be approximated well by only 9 spherical harmonics coefficients per color channel [25]. Thus, to calibrate against a reflectance map, [18] placed a calibration sphere with the same BRDF as the object of interest in the scene. Barron and Malik [1] obviated the sphere and jointly recovered the reflectance map and the surface with a generative model.

In our discriminative approach, we reconstruct the reflectance map directly from an initial estimate of the surface, which we derive solely from the object silhouette. In particular, we map the input image to a sphere according to our silhouette features. The features define a mapping from a pixel $\mathbf{p}$ to polar coordinates on a unit sphere:

$$\sigma(\mathbf{p}): \Omega \to S^2, \quad \sigma(\mathbf{p}) = \left(\cos^{-1} d'_{\text{rel}}(\mathbf{p}), \beta(\mathbf{p})\right) \tag{8}$$

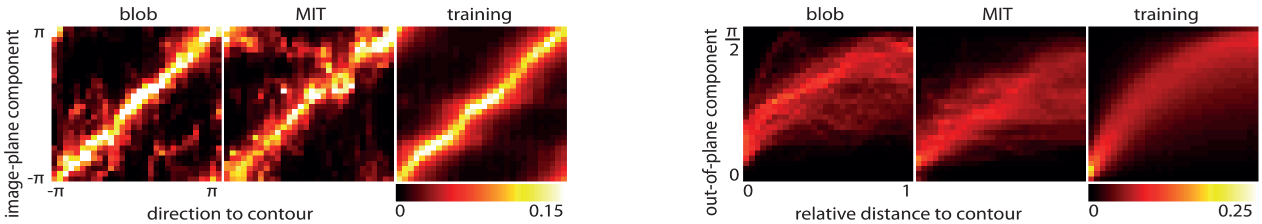

Figure 4. The correlation between pixel position and surface orientation on multiple datasets. In the main columns we plot direction to silhouette *vs.* image-plane component of surface normal and relative distance *vs.* out-of-plane component of normal for 3 datasets each: Blobby shapes [18], MIT intrinsic images [13], and our training dataset from Sec. 4.

We now obtain the color at a polar coordinate (*i.e.*, normal or lighting direction) $\mathbf{s} \in S^2$ by averaging the colors of those input pixels $\mathbf{p}$, whose mapping is a k-nearest neighbor of $\mathbf{s}$:

$$C(\mathbf{s}) = \frac{1}{k}\sum_{i=1}^{k} I(\mathcal{P}_i), \quad \mathcal{P} = \{\mathbf{p} \,|\, \sigma(\mathbf{p}) \in k\text{-NN}(\mathbf{s})\}, \tag{9}$$

where I is the observed color. The number of neighbors k considered is adjusted for the size of the object.

The silhouette features alone give only a coarse estimate of the surface normals. Moreover, certain objects do not fulfill our assumption of being composed of convex parts. A bowl seen from above will cause problems, for example, but likely also for other algorithms that estimate shape and reflectance. Most objects, however, contain limited concavities whose effect on the estimated reflectance map is generally compensated by other convexities.

However, since the mapping from pixels to polar coordinates is many-to-one, we average the colors of points with similar distance and direction to the silhouette. This acts as a low-pass filter, effectively reducing estimation errors from incorrectly mapped points. We thus obtain a robust approximation of a calibration sphere *without* actually having one. From this we can recover the spherical harmonics coefficients of the reflectance in closed form. We found that adjusting the mean and standard deviation of the reflectance map to match the input image (effectively matching brightness and contrast) improves the final estimate.

8. Experiments

8.1. Feature evaluation

To understand the contribution of the various input features, we first analyze the qualitative (Fig. 5) and quantitative impact on surface normal prediction. We evaluated our unary input features on the training set of the MIT intrinsic image dataset, rendered under all illuminations from [18]. Note that these illuminations stem from real environment maps and also contain nearly white illumination; this is in contrast to [1], which re-rendered the MIT data with illuminations sampled from their learned prior. After rendering, we added Gaussian noise ($\sigma = 0.001$) to the images and thresholded values below 0 and above 1. We trained on the dataset described in Sec. 4. Table 1(a) shows the results evaluated using the median angular error (MAE) and the mean-squared error of the normal (nMSE, see [1]).

We use the basic color feature (RGB) as baseline. Adding our silhouette-based features (+Silh) increases the overall performance. Nonetheless, they are better suited for objects that are round or composed of convex parts with a curved surface. Thus, the performance increases significantly on objects fulfilling these assumptions, but only marginally on planar objects or when self-occlusions are present. In colorful illuminations (Fig. 5, bottom), the silhouette features can be misleading in parts, but overall clearly improve performance. Texton features (+Tex) work particularly well under chromatic illumination, indicating that the captured spatial information eliminates many ambiguities; yet even in white illumination they yield a clear benefit. Their combination (+Silh+Tex) works best overall and is robust w.r.t. the illumination conditions.

8.2. Integrability

We evaluate several approaches for enforcing integrability of the estimated surface in Tab. 1(b). As a simple starting point we choose an l_2-penalty on violations of Eq. (3). Next, we employ an l_1-penalty [26], and finally an l_2-penalty under perspective projection following [24]. The unary predictions are our baseline. Without any post-processing, the surface normals can be reconstructed already with good accuracy. Under synthetic illumination, the performance may even decrease when enforcing integrability. We observed, however, that for real images, which potentially violate the Lambertian assumptions, integrability is particularly helpful. Since the objects in the MIT dataset were presumably recorded with a long focal length, the benefits of a perspective approach are negligible. Leveraging the high performance levels of the regression forest, we adapted the l_2-penalty to restrict surface normals to a convex combination of samples drawn from the distributions in the leaf nodes of the trees. This version (*conv*) clearly outperformed all other approaches at the price of a much higher run-time. In further experiments, we thus rely on the simple l_2-penalty.

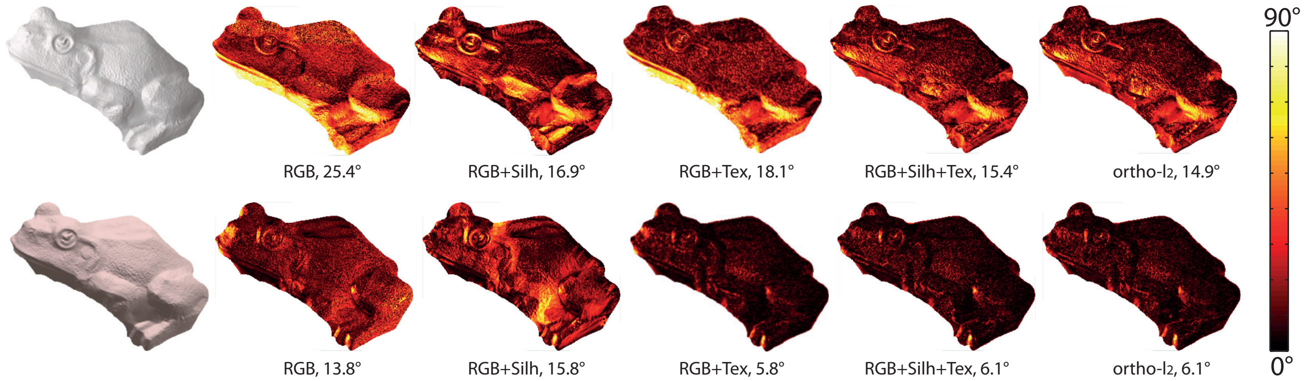

Figure 5. Importance of unary features. For the images in the first column (white illumination – top, colored – bottom), we estimate surfaces using only subsets of features; see text for details. The remaining columns depict the angular error per pixel and its median below.

Features	MAE	nMSE
RGB	13.77°	0.179
RGB+Silh	10.90°	0.130
RGB+Tex	7.92°	0.097
RGB+Silh+Tex	**7.09°**	**0.069**

(a) Results for unary features.

Method	MAE	nMSE	run-time
no integrability	7.09°	0.069	89.0s
l_2, orthographic	7.33°	0.057	98.5s
l_2, orthographic, conv.	**6.46°**	**0.056**	1172.8s
l_1, orthographic	7.34°	0.058	98.0s
l_2, perspective	7.42°	0.059	**97.2s**

(b) Results for enforcing integrability.

Table 1. Influence of unary features and integrability constraints. The run-times includes training, inference, and post-processing.

Method	nMSE*	nMSE	lMSE
Cross scale	0.058	0.471	0.039
Ours	**0.034**	**0.196**	**0.013**

(a) Results on synthetic images (MIT intrinsic [1]).

Illumination	lab[33]	natural (ours)		
Method	MAE	MAE*	MAE	lMSE
Local context	17.27°	–	–	–
Cross scale	19.30°	**20.29°**	29.29°	0.013
Ours	**15.96°**	20.51°	**23.07°**	**0.002**

(b) Results on real images.

Table 2. Comparison to other methods. See text for further explanation. * indicates that the illumination was given.

8.3. Comparison with other methods

We quantitatively compare against 2 state-of-the-art methods on 3 different datasets, 2 of which are provided with the respective methods, and one recorded by ourselves.

The first method we compare to is the shape-from-shading component of the SIRFS method from Barron and Malik [1] (termed "Cross scale"). We use source code provided by the authors and evaluate both under unknown and given illumination. In unknown illumination, we also consider the accuracy of the estimated reflectance map (lMSE, see [1]). Tab. 2(a) gives results on the dataset of [1], a variant of the MIT intrinsic image dataset [13] re-rendered under chromatic illumination.

The second baseline, the method of Xiong *et al.* [33] (termed "Local context") relies on local shading context to infer shape from shading under known illumination. With their method comes a dataset of 10 objects recorded under white directional illumination. Since they also evaluated the shape-from-shading component from [1], we simply restate the results from their paper (Tab. 2(b), first column) and run our algorithm on their dataset.

We did not train our algorithm on any of these datasets, but use our own separate set of artist-created models as described in Sec 4. We used the training split of the MIT intrinsic images once to set the hyperparameters (number of trees, maximum tree depth, *etc.*) with Bayesian optimization [29] and used these settings in all of our experiments.

Real-world experiment. Good performance on synthetic data does not always translate to realistic settings [9]. We aim to address this by quantitatively evaluating our method on real-world data. Unfortunately, no shape-from-shading dataset captured under natural illumination exists so far; methods considering natural illumination were instead evaluated only qualitatively or on synthetic data [1, 18].

To record ground truth data with high accuracy, photometric stereo methods are well established [33]. However, the lighting environment must be carefully designed and controlled, and only a normal map is obtained. Since we require the shape to precisely align with the captured images and also aim to record in natural illumination, we would need to either synthesize the illumination in the lab, violating the real-world assumption, or build a controlled light

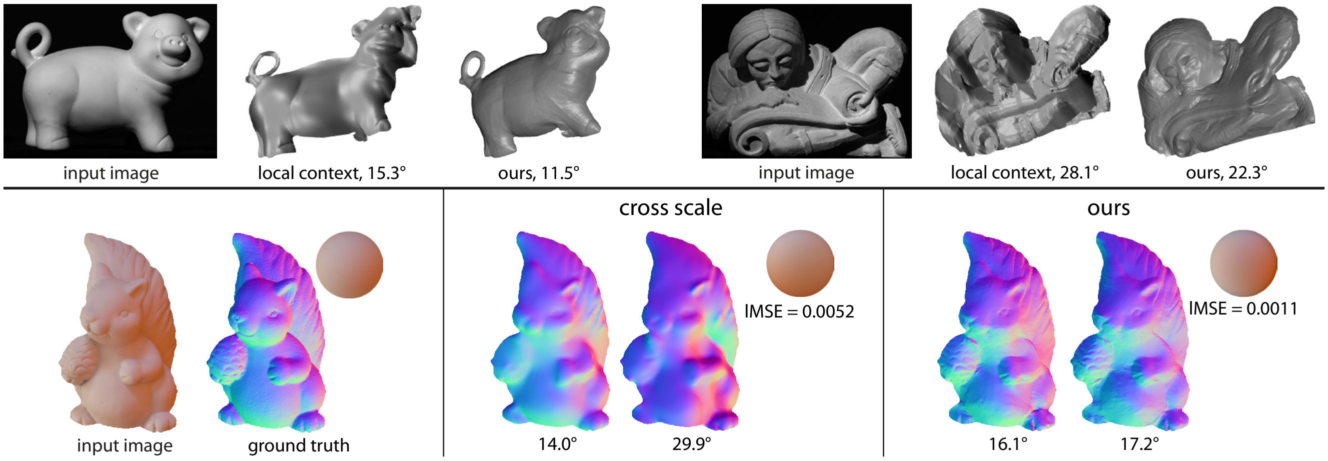

Figure 6. Comparison on real images with median angular errors. For laboratory illumination (top row), we show a novel view of our best and worst result of a reconstructed surface. The input image and the view of "local context" are taken from [33]. For natural illumination (bottom row), we show the surface normal estimates for known (left) and unknown illumination (right), and the estimated reflectance map for the latter case. Across all conditions, our method reconstructs fine surface detail better than previous approaches.

environment at each scene, which in many realistic scenes is next to impossible. Instead, we took ~ 200 pictures for each of 4 objects and reconstructed surface meshes using multi-view stereo [11, 12]; we then aligned the meshes with the test images using mutual information [6]. For the test images taken under real illumination, we painted the objects with a white diffuse paint and recorded them together with a calibration sphere in different environments; latter allows recovering the ground truth illumination. Images and ground truth are publicly available on our website.

We give quantitative results on the dataset in the 3 rightmost columns of Tab 2(b); the reconstructed surfaces are shown in Fig. 6. As before, our method was not specifically adapted to the dataset. The shape prior used by [1] is neither; the shapes used for training (MIT dataset) are still representative (*i.e.*, of similar kind).

We show additional results in Fig. 1 and in the supplementary material.

Results. The quantitative and qualitative results show that our method robustly recovers surfaces and reflectance maps in synthetic, laboratory, and natural illumination. We clearly outperform the cross-scale approach [1] in all metrics; the only exception are the real images with known illumination, where we perform about the same. In the more challenging setting of unknown illumination, we perform significantly better, however. As can be seen in the bottom row of Fig. 6, correctly estimating the reflectance map is crucial to the performance of surface reconstruction. The silhouette features allow for a robust estimate, which is leveraged by our discriminative learning approach. The results in Fig. 6 further highlight that our approach is able to recover fine surface detail on real data, since it does not need to rely on strong spatial regularizers.

We also outperform the local context approach of [33]. One point to note is that our approach can deal with images of different scale (the images of Fig. 6, top are approximately twice the size of those in Fig. 6, bottom). This is due to the scale-invariant nature of our silhouette features.

It may seem surprising that the performance of all methods decreases in realistic settings, since the illumination is more colorful. However, this can be explained by observing that the real data exhibits shadows and fine surface detail, which the synthetic datasets do not. Despite these challenges our discriminative approach is able to provide high-quality surface estimates in uncalibrated illumination.

9. Conclusion

We presented a discriminative learning approach for estimating the shape of an unknown diffuse object with uniform albedo under uncontrolled illumination, given only a single image. We adapted regression forests to predicting surface normals, and proposed and analyzed suitable features that provide local and scale-invariant object-level context without the need for spatial regularization. Pixel-independent predictions are fused by only enforcing integrability of the reconstructed surface. Silhouette features further enable estimating the unknown reflectance map. As with other learning approaches, we need to train our model for each lighting condition. This poses no major drawback, as the combined training and test time of our efficient approach is on par with the test time of other recent methods that do not use learning. We trained our model on novel, large scale training data and evaluated it on several challenging datasets, where it outperforms recent approaches from the literature. Experiments on a new real-world dataset demonstrate its ability to recover fine surface detail outside of the laboratory.

Acknowledgement. We thank Simon Fuhrmann for assistance with multi-view stereo capturing. SRR was supported by the German Research Foundation (DFG) within the GRK 1362. SR was supported in part by the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement No. 307942, as well as by the EU FP7 project "Harvest4D" (No. 323567).

References

[1] J. T. Barron and J. Malik. Color constancy, intrinsic images, and shape estimation. In *ECCV*, 2012. 2, 5, 6, 7, 8

[2] J. T. Barron and J. Malik. Shape, albedo, and illumination from a single image of an unknown object. In *CVPR*, 2012. 2, 3

[3] J. Ben-Arie and D. Nandy. A neural network approach for reconstructing surface shape from shading. In *ICIP*, 1998. 2, 3

[4] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. 3, 4

[5] F. Cole, P. Isola, W. T. Freeman, F. Durand, and E. H. Adelson. ShapeCollage: Occlusion-aware, example-based shape interpretation. In *ECCV*, 2012. 2, 3, 4

[6] M. Cosini, M. Dellepiane, F. Ponchio, and R. Scopigno. Image-to-geometry registration: A mutual information method exploiting illumination-related geometric properties. *Computer Graphics Forum*, 28(7):1755–1764, 2009. 8

[7] I. S. Dhillon and S. Sra. Modeling data using directional distributions. Technical report, TR-03-06, Department of Computer Sciences, The University of Texas at Austin, 2003. 3

[8] Dosch Design. `http://www.doschdesign.com/products/3d/Comic_Characters_V2.html`. 3

[9] J.-D. Durou, M. Falcone, and M. Sagona. Numerical methods for shape-from-shading: A new survey with benchmarks. *CVIU*, 109(1):22–43, 2008. 2, 7

[10] R. Fisher. Dispersion on a sphere. *P. Roy. Soc. Lond. B*, 217(1130), 1953. 3

[11] S. Fuhrmann and M. Goesele. Floating scale reconstruction. In *SIGGRAPH*, 2014. 8

[12] S. Fuhrmann, F. Langguth, and M. Goesele. MVE - A multi-view reconstruction environment. In *GCH*, 2014. 8

[13] R. Grosse, M. K. Johnson, E. H. Adelson, and W. T. Freeman. Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *ICCV*, 2009. 2, 5, 6, 7

[14] A. Hertzmann and S. M. Seitz. Shape and materials by example: A photometric stereo approach. In *CVPR*, 2003. 2

[15] R. Huang and W. A. P. Smith. Shape-from-shading under complex natural illumination. In *ICIP*, 2011. 2

[16] K. Ikeuchi and B. K. Horn. Numerical shape from shading and occluding boundaries. *Artificial Intelligence*, 17(1):141–184, 1981. 1, 5

[17] J. Jancsary, S. Nowozin, and C. Rother. Loss-specific training of non-parametric image restoration models: A new state of the art. In *ECCV*, 2012. 3, 4

[18] M. K. Johnson and E. H. Adelson. Shape estimation in natural illumination. In *CVPR*, 2011. 1, 2, 3, 4, 5, 6, 7

[19] N. Khan, L. Tran, and M. Tappen. Training many-parameter shape-from-shading models using a surface database. In *ICCV Workshops*, 2009. 2, 3, 4

[20] J. J. Koenderink. What does the occluding boundary tell us about solid shape. *Perception*, 13(3), 1984. 1, 5

[21] P. Kontschieder, P. Kohli, J. Shotton, and A. Criminisi. GeoF: Geodesic forests for learning coupled predictors. In *CVPR*, 2013. 2, 3, 4

[22] G. Oxholm and K. Nishino. Shape and reflectance from natural illumination. In *ECCV*, 2012. 2, 5

[23] A. Panagopoulos, S. Hadap, and D. Samaras. Reconstructing shape from dictionaries of shading primitives. In *ACCV*, 2012. 2

[24] T. Papadhimitri and P. Favaro. A new perspective on uncalibrated photometric stereo. In *CVPR*, 2013. 4, 6

[25] R. Ramamoorthi and P. Hanrahan. An efficient representation for irradiance environment maps. In *SIGGRAPH*, 2001. 1, 4, 5

[26] D. Reddy, A. Agrawal, and R. Chellappa. Enforcing integrability by error correction using l_1-minimization. In *CVPR*, 2009. 4, 6

[27] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *CVPR*, 2011. 3

[28] J. Shotton, J. M. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *IJCV*, 81(1):2–23, 2009. 1, 4

[29] J. Snoek, H. Larochelle, and R. P. Adams. Practical Bayesian optimization of machine learning algorithms. In *NIPS*, 2012. 7

[30] D. Stockman. Vienna 2010 35. Flickr, 2010. licensed under `https://creativecommons.org/licenses/by-sa/2.0/`. 1

[31] G.-Q. Wei and G. Hirzinger. Learning shape from shading by a multilayer network. *T. Neural Netw.*, 7(4):985–995, 1996. 2, 3

[32] R. White and D. Forsyth. Combining cues: Shape from shading and texture. In *CVPR*, 2006. 4

[33] Y. Xiong, A. Chakrabarti, R. Basri, S. J. Gortler, D. W. Jacobs, and T. Zickler. From shading to local shape. *TPAMI*, 37(1):67–79, 2014. 1, 7, 8

[34] R. Zhang, P.-S. Tsai, J. E. Cryer, and M. Shah. Shape from shading: A survey. *TPAMI*, 21(8):690–706, 1999. 2