

# Dense Multi-view 3D-reconstruction Without Dense Correspondences

Yvain QUÉAU<sup>1</sup>, Jean MÉLOU<sup>2,3</sup>, Jean-Denis DUROU<sup>2</sup>, and Daniel CREMERS<sup>1</sup>

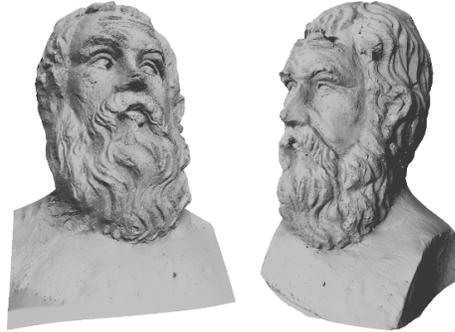
<sup>1</sup>Technical University of Munich, Germany

<sup>2</sup>University of Toulouse, France

<sup>3</sup>Mikros Image, Levallois-Perret, France



$N$  multi-view input real images with illumination



Shape-from-shading without any regularization nor initial estimate

Figure 1: We show how to solve shape-from-shading under natural illumination without regularization, to capture the finest level of detail. When sparse correspondences across multi-view images are available (here, we used  $N = 4$  images from the “Socrates” dataset [48]), unambiguous 3D-reconstruction is achieved. The difficulty of dense matching is thus circumvented.

## Abstract

*We introduce a variational method for multi-view shape-from-shading under natural illumination. The key idea is to couple PDE-based solutions for single-image based shape-from-shading problems across multiple images and multiple color channels by means of a variational formulation. Rather than alternatingly solving the individual SFS problems and optimizing the consistency across images and channels which is known to lead to suboptimal results, we propose an efficient solution of the coupled problem by means of an ADMM algorithm. In numerous experiments on both simulated and real imagery, we demonstrate that the proposed fusion of multiple-view reconstruction and shape-from-shading provides highly accurate dense reconstructions without the need to compute dense correspondences. With the proposed variational integration across multiple views shape-from-shading techniques become applicable to challenging real-world reconstruction problems, giving rise to highly detailed geometry even in areas of smooth brightness variation and lacking texture.*

## 1. Introduction

### 1.1. Multi-view Shape-from-shading

Over the decades the reconstruction of dense 3D geometry from images has been tackled in numerous ways. Two of the most popular strategies are the reconstruction from multiple views using the notion of color or feature correspondence and the reconstruction of shaded objects using the technique of shape-from-shading. Both approaches are in many ways complementary, both have their strengths and limitations. While the fusion of these complementary concepts in a single reconstruction algorithm bears great promise, to date this challenge has remained unsolved and convincing experimental realizations have remained elusive. In this work, we will review existing efforts and propose a novel solution to this challenge.

### 1.2. Related Work

**Multi-view stereo reconstruction.** Multi-view stereo reconstruction (MVS) [15] is among the most powerful techniques to recover 3D geometry from multiple real-world images. The key idea is to exploit the fact that 3D

points are likely to be on the (Lambertian) object surface if the projection into various cameras gives rise to a consistent color, patch or feature value. The arising photo-consistency-weighted minimal surface problems can be optimized using techniques such as graph cuts [44] or convex relaxation [28]. Despite its enormous popularity for real-world reconstruction, multi-view stereo methods have several well-known shortcomings. Firstly, the estimation of dense correspondences is computationally challenging [43]. Secondly, in the absence of color variations (textureless areas), the color consistency assumption degenerates leading to a need for regularity or smoothness assumptions – the resulting photoconsistency-weighted minimal surface formulations degenerate to Euclidean minimal surface problems which exhibit a shrinking bias that leads to the loss of concavities, indentations and other fine-scale geometric details.

**Shape-from-shading.** In contrast to matching features or colors across images, photometric techniques [1] such as shape-from-shading (SFS) [23, 21] explicitly model the reflectance of the object surface. As a result, the brightness variations observed in a single image provides an indication about variations in the normal and geometry. SFS is a classical ill-posed problem with well-known ambiguities such as the one shown in Figure 2. From a single greylevel image both the indentation (red curve) and the protrusion (blue curve) are possible geometric configurations. There exist two main strategies for solving this ambiguity [13, 47]. Variational methods [20] employ regularization. As a result, they provide an approximate SFS solution which is often over-smoothed. Alternatively, methods based on the exact resolution of a nonlinear PDE [30] yield the highest level of detail while implicitly enforcing smoothness in the sense of viscosity solutions. Unfortunately, these PDE solutions lack robustness and they require a boundary condition. Since most shape-from-shading methods require a highly controlled illumination, they often fail when deployed in real-world conditions outside the lab. As shown in Figure 3, existing methods for shape-from-shading under natural illumination [2, 35] strongly depend on the use of a regularization mechanism, which limits their accuracy.



Figure 2: Shape-from-shading suffers from the concave/convex ambiguity (left). We introduce a practical approach to SFS under natural illumination, which achieves unambiguous 3D-reconstruction when sparse correspondences between multi-view images are available (right).

**Shading-based geometry refinement.** Obviously the mentioned concave/convex ambiguity disappears when using more than one observation – see Figure 2, right side. The natural question is therefore how to combine multi-view reconstruction with the concept of shape-from-shading. This has long been identified as a promising track [4], and theoretical guarantees on uniqueness exist [8]. Still, there is a lack of practical multi-view shape-from-shading methods. Jin *et al.* presented in [25] a variational solution, which relies on regularization and may thus miss thin structures. Besides, this solution assumes a single, infinitely distant light source and thus cannot be applied under natural illumination. Methods combining stereo and shading information have also been developed [16, 27, 29, 32, 33, 41, 45, 48]. Yet, they do not fully exploit the potential of shading, because they all consider photometry as a way to refine multi-view 3D-reconstruction, which remains the baseline of the process.

### 1.3. Contribution

In this work, we revisit the challenge of multi-view shape-from-shading. Instead of considering SFS as a post-processing for fine-scale geometric refinement, we rather place it at the core of the multi-view 3D-reconstruction process. The key idea is to model the brightness variations of each color channel and each image by means of a partial differential equation and to subsequently couple these PDE solutions across all images and channels by means of a variational approach. Furthermore, rather than alternately solving for shape-from-shading and consistency across all images (which is known to lead to suboptimal solutions of poor quality), we make use of an efficient ADMM algorithm in order to solve the nonlinearly coupled optimization problem. In numerous experiments we demonstrate that the proposed variational fusion of shape-from-shading across multiple views gives rise to highly accurate dense reconstructions of real-world objects without the need for dense correspondence. We believe that the proposed extension of SFS to multiple views will help to finally bring SFS strategies from the lab into the real world.

### 1.4. Problem Statement and Paper Organization

Given a set of  $N$  input images  $I_i$ ,  $i \in \{1, \dots, N\}$  and the reflectance function  $\mathcal{R}$ , we ultimately wish to estimate  $N$  depth maps  $z_i$ ,  $i \in \{1, \dots, N\}$ , which are both consistent with the observed images (photometric constraint), and consistent with each other (geometric constraint). The proposed framework is of the variational form, and can be written as follows:

$$\min_{\{z_i\}_i} \sum_{i=1}^n \mathcal{P}(\mathcal{R}(z_i) - I_i) + \sum_{1 \leq i < j \leq N} \mathcal{G}(z_i, z_j), \quad (1)$$

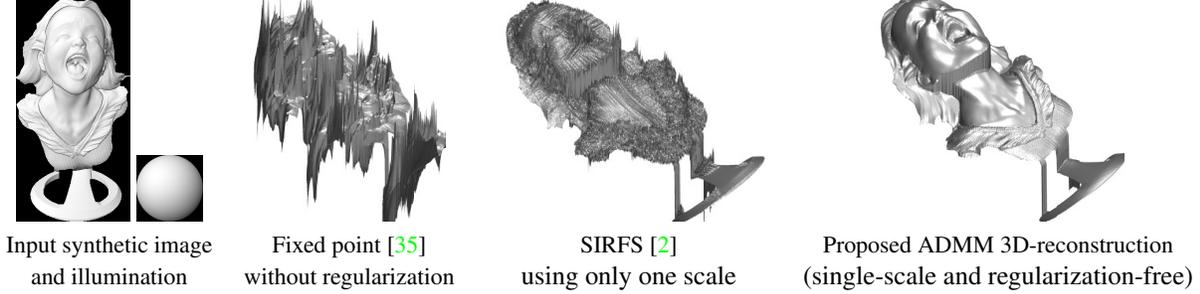


Figure 3: Greylevel shape-from-shading using first-order spherical harmonics. Linearization strategies such as the fixed point one used in [35] induce artifacts if regularization is not employed. Similar issues arise in SIRFS [2] when the multi-scale approach is not used. On the contrary, the proposed ADMM approach provides satisfactory results without resorting to neither of these ad-hoc fixes. In the three experiments, the same initial shape was used (the realistic initialization of Figure 5).

where the photometric energy  $\mathcal{P}$  and the geometric one  $\mathcal{G}$  have to be chosen appropriately in order to ensure that: i) the finest details are being captured; ii) natural illumination can be considered; iii) the solution is not over-smoothed; iv) the  $N$  depth maps are consistent.

The choice of the photometric energy  $\mathcal{P}$  is first discussed in detail in Section 2. It introduces a new approach to SFS under natural illumination which is both variational and PDE-based. It captures the finest details of a surface by avoiding regularization. Yet, since we also avoid using any boundary condition, 3D-reconstruction remains ambiguous if no initial estimate is available. To tackle this issue, we show in Section 3 that sparse correspondences across multi-view images disambiguate the problem.

## 2. Variational SFS Under Natural Illumination

This section introduces an algorithm for solving SFS under general lighting, modeled by channel-dependent, second-order spherical harmonics. We make the same assumptions as in [26] *i.e.*, the lighting and the albedo of the surface are known. In practice, this means that a calibration object (*e.g.*, a sphere) with known geometry and same albedo as the surface to reconstruct must be inserted in the scene. These assumptions are usual in the SFS literature. They could be relaxed by simultaneously estimating shape, illumination and reflectance [2], but we leave this as future work. Instead, we wish to solve SFS without resorting to any prior except differentiability of the depth map.

Our approach relies on the new differential SFS model (5). To solve it in practice, we introduce the variational reformulation (9), which separates the difficulties due to nonlinearity from those due to the non-local nature of the problem. Experimental validation is eventually conducted through an application to shading-based depth refinement.

### 2.1. Image Formation Model and Related Work

Let  $I : \Omega \subset \mathbb{R}^2 \rightarrow \mathbb{R}^C$ ,  $(x, y) \mapsto I(x, y) = [I^1(x, y), \dots, I^C(x, y)]^\top$ , be a greylevel ( $C = 1$ ) or multi-channel ( $C > 1$ ) image of a surface, where  $\Omega$  represents a “mask” of the object being pictured.

We assume that the surface is Lambertian, so its reflectance is characterized by the albedo  $\rho$ . We further consider a second-order spherical harmonics model [3, 38] for the lighting  $\mathbf{l}$ . To deal with the spectral dependencies of reflectance and lighting, we assume both  $\rho$  and  $\mathbf{l}$  are channel-dependent. The albedo is thus a function  $\rho : \Omega \rightarrow \mathbb{R}^C$ ,  $(x, y) \mapsto \rho(x, y) = [\rho^1(x, y), \dots, \rho^C(x, y)]^\top$ , and the lighting in each channel  $c \in \{1, \dots, C\}$  is represented as a vector  $\mathbf{l}^c = [l_1^c, l_2^c, l_3^c, l_4^c, l_5^c, l_6^c, l_7^c, l_8^c, l_9^c]^\top \in \mathbb{R}^9$ .

Eventually, let  $\mathbf{n} : \Omega \rightarrow \mathbb{S}^2 \subset \mathbb{R}^3$ ,  $(x, y) \mapsto \mathbf{n}(x, y) = [n_1(x, y), n_2(x, y), n_3(x, y)]^\top$  be the field of unit-length outward normals to the surface.

With these notations, the image value in each channel  $c \in \{1, \dots, C\}$  writes as follows,  $\forall (x, y) \in \Omega$ :

$$I^c(x, y) = \rho^c(x, y) \mathbf{l}^c \cdot \begin{bmatrix} \mathbf{n}(x, y) \\ 1 \\ n_1(x, y)n_2(x, y) \\ n_1(x, y)n_3(x, y) \\ n_2(x, y)n_3(x, y) \\ n_1(x, y)^2 - n_2(x, y)^2 \\ 3n_3(x, y)^2 - 1 \end{bmatrix}. \quad (2)$$

Our goal is to recover the object shape, given its image, its albedo and the lighting. Each unit-length normal vector  $\mathbf{n}(x, y)$  has two degrees of freedom, thus it is in general impossible to solve Equation (2) independently in each pixel  $(x, y)$ . In particular, if  $C = 1$  and lighting is directional ( $l_4^c = \dots = l_9^c = 0$ ), Equation (2) is a single scalar equation with two unknowns. This particular situation characterizes the classic SFS problem, which is ill-posed [21]. Its resolution has given rise to a number of methods [13, 47].

Yet, few SFS methods deal with non-directional lighting. Near-field pointwise lighting has been shown to help resolving the ambiguities [37], but only partly [6]. Besides, to deal with more diffuse lighting such as natural outdoor illumination, spherical harmonics are better suited. First-order harmonics have been considered in [22], but they only capture up to 90% of “real” lighting, while this rate is over 99% using second-order harmonics [14].

In the context of SFS, second-order harmonics have been used in [2, 26, 39]. The SFS approach of Johnson and Adelson [26] has the same objective as ours *i.e.*, handling multi-channel images and “natural” illumination, knowing the albedo and the lighting. It is shown that this general illumination model actually limits the ambiguities of SFS, since it is the intermediate case between SFS and color photometric stereo [19]. However, this work relies on regularization terms, which favors over-smoothed surfaces. Barron and Malik solve in [2] the more challenging problem of shape, illumination and reflectance from shading (SIRFS). By fixing the albedo and the lighting, and removing all the regularization terms, SIRFS can be applied to SFS. However, the proposed method “fails badly” [2] if a multi-scale strategy is not considered (see Figure 3). Let us also mention for completeness the recent work in [39], which has similar goals as ours (shape-from-shading under natural illumination), but follows an entirely different track based on discriminative learning, which requires prior training.

Overall, there exists no purely data-driven approach to SFS under natural illumination. The rest of this section aims at filling this gap.

## 2.2. Differential Model

Since Equation (2) cannot be solved locally, it must be solved *globally* over the entire domain  $\Omega$ . This can be achieved by assuming surface smoothness. However, in order to prevent losing the fine-scale surface details, this assumption should be as minimal as possible. In particular, regularization terms, which have been widely explored in early SFS works [20, 23], may over-smooth the solution. Instead of having the normal vectors as unknowns and penalizing their variations, as achieved for instance in [26], we rather directly estimate the underlying depth map. To this end, we resort to a differential approach building upon PDEs [30]. This has the advantage of implicitly enforcing differentiability without requiring any regularization term. Let us thus first rewrite (2) as a PDE.

Let the shape be represented as a function  $z : \Omega \rightarrow \mathbb{R}$ , which is the depth map under orthographic projection, and the log of the depth map under perspective projection. In both cases, the normal to the surface is given by

$$\mathbf{n} = \frac{1}{d(z_x, z_y)} \begin{bmatrix} fz_x \\ fz_y \\ -1 - \tilde{x}z_x - \tilde{y}z_y \end{bmatrix}, \quad (3)$$

where:  $\nabla z = [z_x, z_y]^\top$  is the gradient of  $z$ ;  $(f, \tilde{x}, \tilde{y}) = (1, 0, 0)$  under orthographic projection while, under perspective projection,  $f$  is the focal length and  $(\tilde{x}, \tilde{y}) = (x - x_0, y - y_0)$ , with  $(x_0, y_0)$  the coordinates of the principal point; and  $d(z_x, z_y)$  is a coefficient of normalization:

$$d(z_x, z_y) = \sqrt{(fz_x)^2 + (fz_y)^2 + (1 + \tilde{x}z_x + \tilde{y}z_y)^2}. \quad (4)$$

Plugging (3) into (2), we obtain,  $\forall c \in \{1, \dots, C\}$ , the following nonlinear PDE in the depth  $z$  over  $\Omega$ :

$$\mathbf{a}^c(z_x, z_y) \cdot \begin{bmatrix} z_x \\ z_y \end{bmatrix} = b^c(z_x, z_y), \quad (5)$$

with the following definitions for the fields  $\mathbf{a}^c(z_x, z_y) : \Omega \rightarrow \mathbb{R}^2$  and  $b^c(z_x, z_y) : \Omega \rightarrow \mathbb{R}$ :

$$\mathbf{a}^c(z_x, z_y) = \frac{\rho^c}{d(z_x, z_y)} \begin{bmatrix} fl_1^c - \tilde{x}l_3^c \\ fl_2^c - \tilde{y}l_3^c \end{bmatrix}, \quad (6)$$

$$b^c(z_x, z_y) = I^c - \rho^c \begin{bmatrix} l_3^c \\ l_4^c \\ l_5^c \\ l_6^c \\ l_7^c \\ l_8^c \\ l_9^c \end{bmatrix} \cdot \begin{bmatrix} \frac{-1}{d(z_x, z_y)} \\ 1 \\ \frac{f^2 z_x z_y}{d(z_x, z_y)^2} \\ \frac{f z_x (-1 - \tilde{x}z_x - \tilde{y}z_y)}{d(z_x, z_y)^2} \\ \frac{f z_y (-1 - \tilde{x}z_x - \tilde{y}z_y)}{d(z_x, z_y)^2} \\ \frac{f^2 (z_x^2 - z_y^2)}{d(z_x, z_y)^2} \\ \frac{3(-1 - \tilde{x}z_x - \tilde{y}z_y)^2}{d(z_x, z_y)^2} - 1 \end{bmatrix}. \quad (7)$$

Various methods have been suggested for solving PDEs akin to (5), in some specific cases. When  $C = 1$ , and lighting is directional and frontal (*i.e.*,  $l_3$  is the only non-zero lighting component), then (5) becomes the *eikonal equation*, which was first exhibited for SFS in [7]. After this inverse problem has caught the attention of several mathematicians [30, 40], efficient numerical methods for approximating solutions to this well-known equation have been suggested, using for instance semi-Lagrangian schemes [12]. Under perspective projection, an eikonal-like equation also arises [36, 42]. The case where lighting is depth-dependent (so-called attenuation factor) is also interesting as it is less ambiguous [37]. Semi-Lagrangian schemes can also be used for the resolution, see for instance [6]. Still, most of these differential methods require a boundary condition, or at least a state constraint, which are rarely available in practice. In addition, there currently lacks a purely data-driven numerical SFS method which would handle second-order lighting and multi-channel images: [2] is strongly dependent on a multi-scale strategy, and [26] is non-differential (per-pixel surface normal estimation) and thus resorts to regularization. The variational approach discussed hereafter solves all these issues at once.

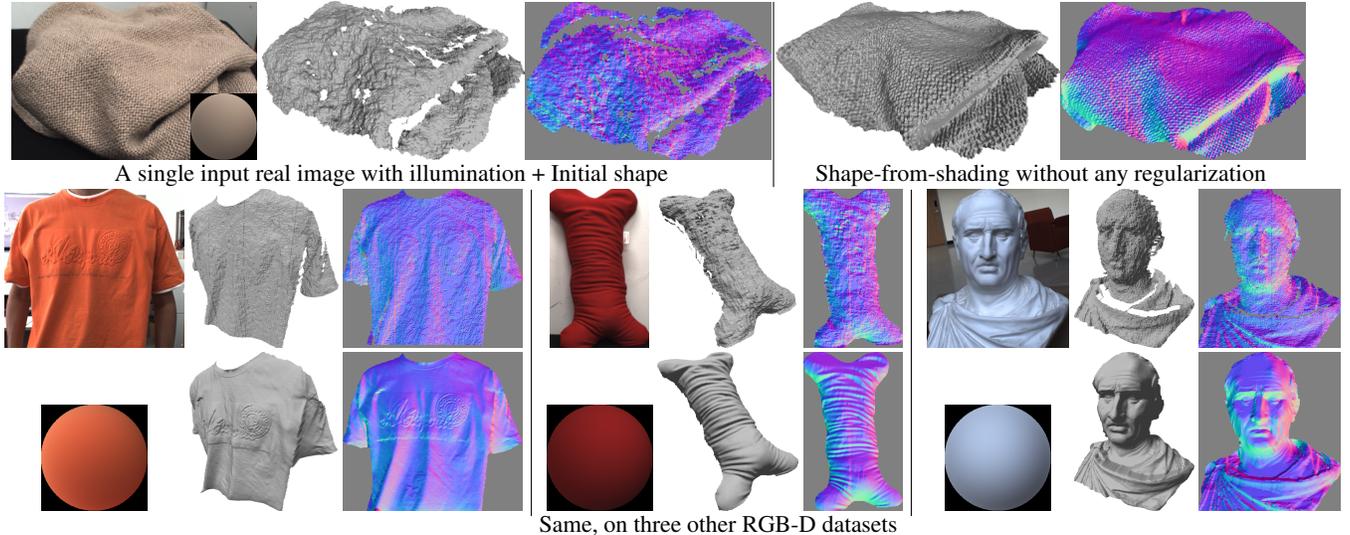


Figure 4: Depth refinement of RGB-D data using the proposed SFS method.

### 2.3. Variational Formulation

The  $C$  PDEs in (5) are in general incompatible due to noise. Thus, an approximate solution must be sought. For simplicity, we follow here a least-squares approach:

$$\min_{z: \Omega \rightarrow \mathbb{R}} \sum_{c=1}^C \left\| \mathbf{a}^c(z_x, z_y) \cdot \begin{bmatrix} z_x \\ z_y \end{bmatrix} - b^c(z_x, z_y) \right\|_2^2, \quad (8)$$

where  $\|\cdot\|_2$  is the  $\ell^2$  norm over the domain  $\Omega$ .

If the fields  $\mathbf{a}^c$  and  $b^c$  were not dependent on  $z$ , then (8) would be a linear least-squares problem. In recent works on shading-based refinement [35], it is suggested to proceed iteratively, by freezing these terms at each iteration. Although this ‘‘fixed point’’ strategy looks appealing, Figure 3 shows that it induces artifacts, and thus regularization must be employed [35]. Other artifacts also arise in SIRFS [2], when the multi-scale strategy is not employed.

Instead of eliminating artifacts by regularization, which may induce a loss of geometric details, we rather separate the difficulty induced by the nonlinearity from that induced by the dependency on the gradient. To this end, we introduce an auxiliary variable  $\theta: \Omega \rightarrow \mathbb{R}^2$ , and rewrite (8) in the following, equivalent, manner:

$$\begin{aligned} \min_{\substack{z: \Omega \rightarrow \mathbb{R} \\ \theta: \Omega \rightarrow \mathbb{R}^2}} \sum_{c=1}^C \left\| \mathbf{a}^c(\theta) \cdot \begin{bmatrix} z_x \\ z_y \end{bmatrix} - b^c(\theta) \right\|_2^2 \\ \text{s.t. } (z_x, z_y) = \theta. \end{aligned} \quad (9)$$

We then turn (9) into a sequence of simpler problems through an ADMM algorithm [5]. The augmented La-

grangian functional associated to (9) is defined as

$$\begin{aligned} \mathcal{L}_\beta(z, \theta, \lambda) = \sum_{c=1}^C \left\| \mathbf{a}^c(\theta) \cdot \begin{bmatrix} z_x \\ z_y \end{bmatrix} - b^c(\theta) \right\|_2^2 \\ + \langle \lambda, (z_x, z_y) - \theta \rangle + \frac{\beta}{2} \|(z_x, z_y) - \theta\|_2^2, \end{aligned} \quad (10)$$

where  $\lambda$  represent Lagrange multipliers, and  $\beta > 0$ . ADMM then minimizes (9) by the following iterations:

$$z^{(k+1)} = \underset{z}{\operatorname{argmin}} \mathcal{L}_{\beta^{(k)}}(z, \theta^{(k)}, \lambda^{(k)}), \quad (11)$$

$$\theta^{(k+1)} = \underset{\theta}{\operatorname{argmin}} \mathcal{L}_{\beta^{(k)}}(z^{(k+1)}, \theta, \lambda^{(k)}), \quad (12)$$

$$\lambda^{(k+1)} = \lambda^{(k)} + \beta^{(k)} \left( (z_x^{(k+1)}, z_y^{(k+1)}) - \theta^{(k+1)} \right). \quad (13)$$

where  $\beta^{(k)}$  is determined automatically [18].

We then discretize (11) by finite differences, and solve the discrete optimality conditions by conjugate gradient. With this approach, no explicit boundary condition is required. As for (12), it is solved in each pixel by a Newton method [11]. In our experiments, the algorithm stops when the relative residual of the energy in (8) falls below  $10^{-3}$ .

### 2.4. Experiments

Since our method estimates a locally optimal solution, initialization matters. There is one situation where a reasonable initial estimate is available. This is when using an RGB-D camera: the depth channel  $D$  is noisy, but it may be refined using shading [9, 10, 17, 34, 35, 46]. Hence, to qualitatively evaluate our approach, we consider in Figure 4 three real-world RGB-D datasets from [17], estimating lighting from the rough depth map (assuming  $\rho \equiv 1$ ). We attain the finest level of surface detail possible, since the surface is not over-smoothed through regularization.

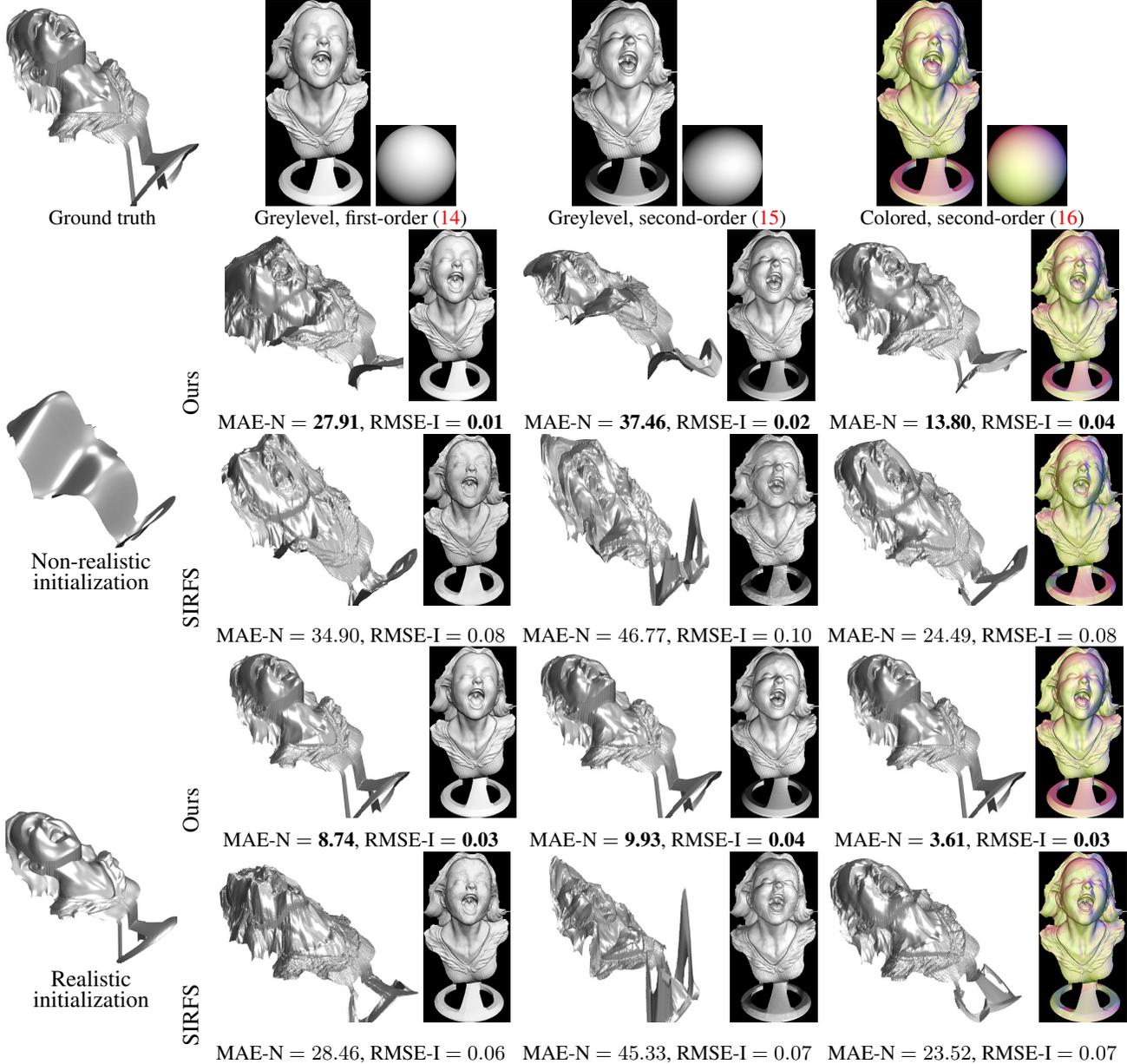


Figure 5: Evaluation of our SFS approach against the multi-scale one from SIRFS [2], in three different lighting situations and using two different initial estimates. For each experiment, we provide the mean angular error w.r.t. ground truth normals (MAE-N, in degrees), and the root mean square error between the input synthetic image and the one simulated from the estimated depth (RMSE-I). Our method outperforms SIRFS in all tests, in terms of both metrics.

For quantitative evaluation, we use in Figure 5 the well-know “Joyful Yell” dataset, using three lighting scenarios. We first consider greylevel images, with a single-order and then a second-order lighting model, respectively defined by:

$$\mathbf{l}^1 = [0.1, -0.25, -0.7, 0.2, 0, 0, 0, 0, 0]^\top, \quad (14)$$

$$\mathbf{l}^2 = [0.2, 0.3, -0.7, 0.5, -0.2, -0.2, 0.3, 0.3, 0.2]^\top. \quad (15)$$

In the third experiment, we consider a colored, second-order lighting model defined by:

$$\mathbf{l}^3 = \begin{bmatrix} -0.2 & -0.2 & -1 & 0.4 & 0.1 & -0.1 & -0.1 & -0.1 & 0.05 \\ 0 & 0.2 & -1 & 0.3 & 0 & 0.2 & 0.1 & 0 & 0.1 \\ 0.2 & -0.2 & -1 & 0.2 & -0.1 & 0 & 0 & 0.1 & 0 \end{bmatrix}^\top. \quad (16)$$

The importance of initialization is assessed by using two different initial estimates. The accuracy of 3D-reconstruction is evaluated by the mean angular error between the recovered normals and the ground truth ones, and the ability to explain the input image is measured through the RMSE between the data and the image simulated from the 3D-reconstruction.

We compared those values against SIRFS [2], which is the only method for SFS under natural illumination whose code is freely available. For fair comparison, we disabled albedo and lighting estimation in SIRFS, and gave a zero weight to all smoothing terms. To avoid the artifacts shown in Figure 3, SIRFS’s multi-scale strategy was used. Figure 5 proves that SFS under natural illumination can be solved using a purely data-driven strategy, without resorting neither to regularization nor to multi-scale. Besides, the runtimes of our method and SIRFS are comparable: a few minutes in all cases, for images with 150.000 non-black pixels.

Still, these experiments show that the proposed method strongly depends on the choice of the initial estimate. We now show how to better constrain the 3D-reconstruction problem through sparse multi-view correspondences.

### 3. Multi-view Shape-from-shading

Although colored natural illumination partly disambiguates SFS, it does not entirely remove ambiguities [26]. Another disambiguation strategy must be considered in the absence of a good initial estimate. We now show that sparse correspondences in a multi-view framework can be employed for this purpose.

To this end, let us now assume that we are given  $N$  images  $\{I_i : \Omega_i \subset \mathbb{R}^2 \rightarrow \mathbb{R}^C\}_{i \in \{1, \dots, N\}}$ , along with the corresponding albedo maps and lighting vectors, both assumed to be channel- and image-dependent and denoted by  $\{\rho_i : \Omega \rightarrow \mathbb{R}^C\}_i$  and  $\{\mathbf{l}_i^c\}_{c,i}$ . The joint resolution of the  $N$  SFS problems could be achieved by solving  $N$  variational problems such as (9). However, this would result in  $N$  inconsistent depth maps: the  $N$  SFS problems need to be coupled.

#### 3.1. Sparse Multi-view Constraints

We use multi-view consistency to couple the  $N$  SFS problems, and show that ambiguities are limited when introducing sparse correspondences between the images. We conjecture that any ambiguity even disappears if the correspondence set is dense. This conjecture could probably be proved by following [8], but we leave this as future work.

Let us assume that some sparse inter-images pixel correspondences are given (which can be obtained, for instance, by matching SIFT descriptors), and let us write them as the following  $\Omega_i \times \Omega_j \rightarrow \mathbb{R}$  functions, where  $\Omega_i$  and  $\Omega_j$  are the

masks of the object in images  $i$  and  $j$ ,  $i < j$ :

$$c_{i,j}(\mathbf{p}_i, \mathbf{p}_j) = \begin{cases} 1 & \text{if pixel } \mathbf{p}_i \text{ in image } I_i \text{ is matched} \\ & \text{with pixel } \mathbf{p}_j \text{ in image } I_j, \\ 0 & \text{otherwise.} \end{cases} \quad (17)$$

Assuming perspective projection, a 3D-point  $\mathbf{x}$  in world coordinates is conjugate to a pixel  $\mathbf{p}_i$  according to

$$\mathbf{x} = e^{z_i(\mathbf{p}_i)} \mathbf{R}_i \left[ \frac{1}{f_i} \tilde{\mathbf{p}}_i^\top, 1 \right]^\top + \mathbf{t}_i, \quad \forall \mathbf{p}_i \in \Omega_i, \quad (18)$$

where  $e^{z_i}$  is the  $i$ -th depth map (recall that we set  $z_i$  to the log depth map under perspective projection),  $\tilde{\mathbf{p}}_i$  is the pixel coordinates w.r.t. the  $i$ -th principal point,  $f_i$  is the  $i$ -th focal length, and  $\mathbf{R}_i \in \mathbb{R}^{3 \times 3}$  and  $\mathbf{t}_i \in \mathbb{R}^3$  are the rotation and translation describing the  $i$ -th pose of the camera (we assume that these poses are calibrated).

The multi-view consistency constraint then writes

$$c_{i,j}(\mathbf{p}_i, \mathbf{p}_j) \left( e^{z_i(\mathbf{p}_i)} \mathbf{R}_i \left[ \frac{1}{f_i} \tilde{\mathbf{p}}_i^\top, 1 \right]^\top - e^{z_j(\mathbf{p}_j)} \mathbf{R}_j \left[ \frac{1}{f_j} \tilde{\mathbf{p}}_j^\top, 1 \right]^\top \right) - c_{i,j}(\mathbf{p}_i, \mathbf{p}_j) (\mathbf{t}_j - \mathbf{t}_i) = \mathbf{0}, \quad (19)$$

which we rewrite as the following nonlinear constraint:

$$\mathbf{C}_{i,j}(z_i, z_j) - \mathbf{d}_{i,j} = 0, \quad (20)$$

where  $\mathbf{C}_{i,j}(z_i, z_j)$  is a  $\Omega_i \times \Omega_j \rightarrow \mathbb{R}^3$  function depending on the depth maps  $z_i$  and  $z_j$ , whereas the function  $\mathbf{d}_{i,j} : \Omega_i \times \Omega_j \rightarrow \mathbb{R}^3$  does not.

#### 3.2. Proposed Variational Paradigm

To disambiguate SFS through multi-views, we suggest to use  $\mathcal{G}(z_i, z_j) = \lambda \|\mathbf{C}_{i,j}(z_i, z_j) - \mathbf{d}_{i,j}\|_{2,i,j}^2$  in the variational model (1), where  $\|\cdot\|_{2,i,j}$  is the  $\ell^2$  norm over  $\Omega_i \times \Omega_j$ , and  $\lambda \geq 0$  is a weighting factor. Since the constraint (20) only depends on the depth values, and not on their gradients, we rather write it in terms of the auxiliary variables of the ADMM algorithm. This is motivated by the fact that the updates of these variables already require per-pixel nonlinear least-squares optimization. Moreover, the depth updates remain linear least-squares ones if the multi-view constraint is written in terms of the auxiliary variables. We thus define new auxiliary variables  $\theta_i = ((z_i)_x, (z_i)_y, z_i)$ , and turn (9) into:

$$\begin{aligned} \min_{\substack{\{z_i: \Omega_i \rightarrow \mathbb{R}\}_i \\ \{\theta_i: \Omega_i \rightarrow \mathbb{R}^3\}_i}} & \sum_{i=1}^N \sum_{c=1}^C \left\| \mathbf{a}^c(\theta_i) \cdot \begin{bmatrix} (z_i)_x \\ (z_i)_y \end{bmatrix} - b^c(\theta_i) \right\|_{2,i}^2 \\ & + \frac{\lambda}{2} \sum_{1 \leq i < j \leq N} \|\mathbf{C}_{i,j}(\theta_i, \theta_j) - \mathbf{d}_{i,j}\|_{2,i,j}^2 \\ \text{s.t.} & ((z_i)_x, (z_i)_y, z_i) = \theta_i, \quad \forall i \in \{1, \dots, N\}. \end{aligned} \quad (21)$$

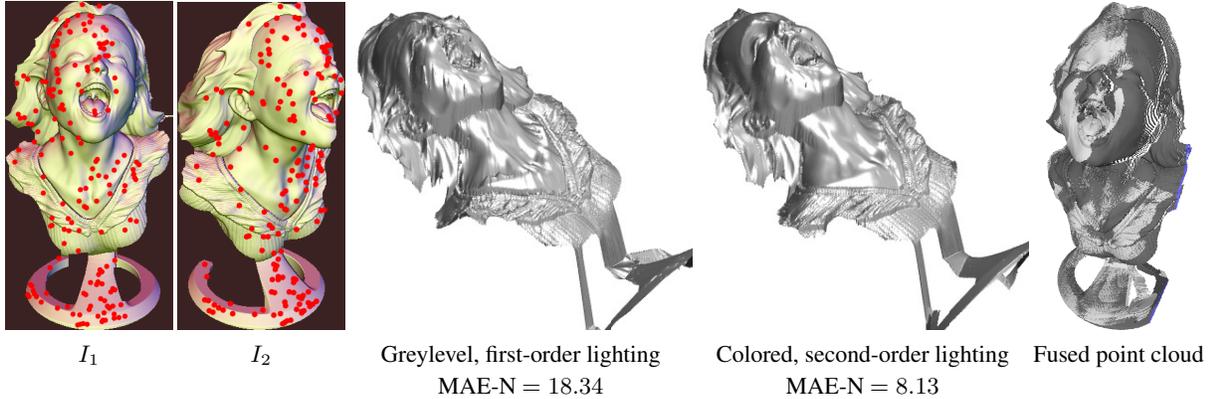


Figure 6: Binocular shape-from-shading. Left: input synthetic images and sparse correspondences, under the same colored, second-order lighting as in Figure 5. With  $N = 2$  views, SFS is disambiguated (no initial estimate is needed) and the 3D-reconstruction error is largely decreased. On the right, we show the point cloud obtained by fusing both depth maps  $z_1$  and  $z_2$  obtained from the color 3D-reconstruction.

We experimentally found that the choice of a particular value of the parameter  $\lambda$  is not important. Obviously, if  $\lambda$  is set to 0, then the  $N$  SFS are uncoupled, and thus ambiguous. Yet, as long as  $\lambda$  is “high enough”, ambiguities disappear. In our tests, we found that the range  $\lambda \in [10^{-8}, 10^{-2}]$  provides comparable results, and always used the value  $\lambda = 10^{-5}$ .

It is straightforward to modify the previous ADMM algorithm for solving (21). In Figure 6, we show the 3D-reconstructions obtained from  $N = 2$  synthetic views, in the same lighting scenarios as in the first and third experiments of Figure 5, using the same non-realistic initial estimate. We used 173 pixel correspondences which were randomly picked using the ground-truth geometry. In comparison with the single-view results (see Figure 5), the estimated depth maps are more accurate. Besides, if we fuse both depth maps into a point cloud (using the known camera poses), we observe that both 3D-reconstructions are “consistent”, which proves that ambiguities are eliminated.

Eventually, we present in Figures 1 and 7 the results of our method on two real-world datasets from [48]. We chose these datasets because they exhibit a uniform, though unknown, albedo. This albedo can thus be estimated during lighting calibration (since illumination is not provided in these datasets, it was calculated from the 3D-reconstructions provided in [48], but these 3D-reconstructions were then not used any further). Sparse correspondences were extracted by matching standard SIFT features [31] (the total number of used matches is worth 7982 for the  $N = 4$  images of the “Sokrates” dataset, and 162 for the  $N = 2$  images of the “Figure” one). These real-world experiments demonstrate that shading-based multi-view 3D-reconstruction constitutes a promising alternative to standard dense multi-view stereo.

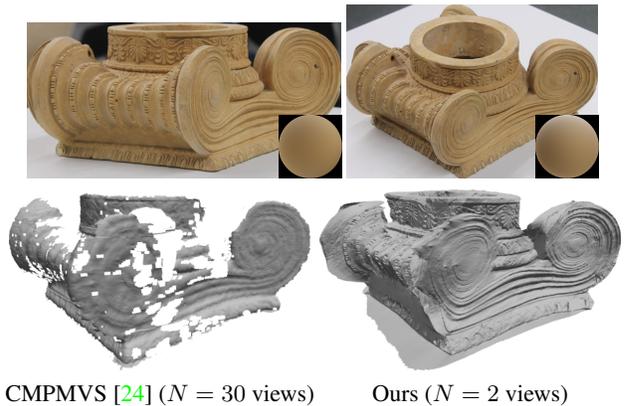


Figure 7: 3D-reconstruction of the “Figure” object [48]. Top:  $N = 2$  input real images  $I_1$  and  $I_2$ . Bottom: depth map  $z_1$  estimated using MVS (left, we show one out of  $N = 30$  depth maps from CPMVS [24]), and the proposed multi-view SFS method (right). The latter yields a much more dense reconstruction with more fine-scale details, although only  $N = 2$  views are used.

#### 4. Conclusion

We have shown how to achieve dense multi-view 3D-reconstruction without dense correspondences. A new variational approach to shape-from-shading under general lighting is used as the main tool for densification. It allows to drastically reduce the number of required images, while improving the amount of detail in the 3D-reconstruction. In future work, the new approach may be extended by automatic estimation of the albedo and of the lighting. This would allow coping with a broader variety of surfaces, and simplify the overall procedure.

## References

- [1] J. Ackermann and M. Goesele. A survey of photometric stereo techniques. *Foundations and Trends in Computer Graphics and Vision*, 9(3-4):149–254, 2015. 2
- [2] J. T. Barron and J. Malik. Shape, illumination, and reflectance from shading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(8):1670–1687, 2015. 2, 3, 4, 5, 6, 7
- [3] R. Basri and D. P. Jacobs. Lambertian reflectances and linear subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(2):218–233, 2003. 3
- [4] A. Blake, A. Zisserman, and G. Knowles. Surface descriptions from stereo and shading. *Image and Vision Computing*, 3(4):183–191, 1985. 2
- [5] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011. 5
- [6] M. Breuß, E. Cristiani, J.-D. Durou, M. Falcone, and O. Vogel. Perspective shape from shading: Ambiguity analysis and numerical approximations. *SIAM Journal on Imaging Sciences*, 5(1):311–342, 2012. 4
- [7] A. R. Bruss. The eikonal equation: Some results applicable to computer vision. *Journal of Mathematical Physics*, 23(5):890–896, 1982. 4
- [8] A. Chambolle. A uniqueness result in the theory of stereo vision: coupling shape from shading and binocular information allows unambiguous depth reconstruction. *Annales de l’IHP - Analyse non linéaire*, 11(1):1–16, 1994. 2, 7
- [9] A. Chatterjee and V. Madhav Govindu. Photometric refinement of depth maps for multi-albedo objects. In *Proceedings of CVPR*, 2015. 5
- [10] G. Choe, J. Park, Y.-W. Tai, and I. S. Kweon. Refining geometry from depth sensors using IR shading images. *International Journal of Computer Vision*, 122(1):1–16, 2017. 5
- [11] T. F. Coleman and Y. Li. An interior trust region approach for nonlinear minimization subject to bounds. *SIAM Journal on Optimization*, 6(2):418–445, 1996. 5
- [12] E. Cristiani and M. Falcone. Fast semi-lagrangian schemes for the eikonal equation and applications. *SIAM Journal on Numerical Analysis*, 45(5):1979–2011, 2007. 4
- [13] J.-D. Durou, M. Falcone, and M. Sagona. Numerical methods for shape-from-shading: A new survey with benchmarks. *Computer Vision and Image Understanding*, 109(1):22–43, 2008. 2, 3
- [14] D. Frolova, D. Simakov, and R. Basri. Accuracy of spherical harmonic approximations for images of lambertian objects under far and near lighting. In *Proceedings of ECCV*, 2004. 4
- [15] Y. Furukawa and C. Hernández. Multi-view stereo: A tutorial. *Foundations and Trends in Computer Graphics and Vision*, 9(1-2):1–148, 2015. 1
- [16] S. Galliani and K. Schindler. Just look at the image: Viewpoint-specific surface normal prediction for improved multi-view reconstruction. In *Proceedings of CVPR*, 2016. 2
- [17] Y. Han, J.-Y. Lee, and I. S. Kweon. High quality shape from a single RGB-D image under uncalibrated natural illumination. In *Proceedings of ICCV*, 2013. 5
- [18] B. S. He, H. Yang, and S. L. Wang. Alternating direction method with self-adaptive penalty parameters for monotone variational inequalities. *Journal of Optimization Theory and Applications*, 106(2):337–356, 2000. 5
- [19] C. Hernández, G. Vogiatzis, G. J. Brostow, B. Stenger, and R. Cipolla. Non-rigid Photometric Stereo with Colored Lights. In *Proceedings of ICCV*, 2007. 4
- [20] B. K. P. Horn and M. J. Brooks. The variational approach to shape from shading. *Computer Vision, Graphics, and Image Processing*, 33(2):174–208, 1986. 2, 4
- [21] B. K. P. Horn and M. J. Brooks, editors. *Shape from Shading*. MIT Press, 1989. 2, 3
- [22] R. Huang and W. A. P. Smith. Shape-from-shading under complex natural illumination. In *Proceedings of ICIP*, 2011. 4
- [23] K. Ikeuchi and B. K. Horn. Numerical shape from shading and occluding boundaries. *Artificial intelligence*, 17(1-3):141–184, 1981. 2, 4
- [24] M. Jancosek and T. Pajdla. Multi-view reconstruction preserving weakly-supported surfaces. In *Proceedings of CVPR*, 2011. 8
- [25] H. Jin, D. Cremers, D. Wang, A. Yezzi, E. Prados, and S. Soatto. 3-D Reconstruction of Shaded Objects from Multiple Images Under Unknown Illumination. *International Journal of Computer Vision*, 76(3):245–256, 2008. 2
- [26] M. K. Johnson and E. H. Adelson. Shape estimation in natural illumination. In *Proceedings of CVPR*, 2011. 3, 4, 7
- [27] K. Kim, A. Torii, and M. Okutomi. Multi-view Inverse Rendering Under Arbitrary Illumination and Albedo. In *Proceedings of ECCV*, 2016. 2
- [28] K. Kolev, M. Klodt, T. Brox, and D. Cremers. Continuous global optimization in multiview 3d reconstruction. *International Journal of Computer Vision*, 2009. 2
- [29] F. Langguth, K. Sunkavalli, S. Hadap, and M. Goesele. Shading-aware Multi-view Stereo. In *Proceedings of ECCV*, 2016. 2
- [30] P.-L. Lions, E. Rouy, and A. Tourin. Shape-from-shading, viscosity solutions and edges. *Numerische Mathematik*, 64(1):323–353, 1993. 2, 4
- [31] D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of ICCV*, 1999. 8
- [32] D. Maurer, Y. C. Ju, M. Breuß, and A. Bruhn. Combining Shape from Shading and Stereo: A Variational Approach for the Joint Estimation of Depth, Illumination and Albedo. In *Proceedings of BMVC*, 2016. 2
- [33] D. Nehab, S. Rusinkiewicz, J. Davis, and R. Ramamoorthi. Efficiently combining positions and normals for precise 3D geometry. *ACM Transactions on Graphics*, 24(3):536–543, 2005. 2
- [34] R. Or-El, R. Hershkovitz, A. Wetzler, G. Rosman, A. M. Bruckstein, and R. Kimmel. Real-time depth refinement for specular objects. In *Proceedings of CVPR*, 2016. 5
- [35] R. Or-El, G. Rosman, A. Wetzler, R. Kimmel, and A. Bruckstein. RGBD-Fusion: Real-Time High Precision Depth Recovery. In *Proceedings of CVPR*, 2015. 2, 3, 5

- [36] E. Prados and O. Faugeras. Perspective shape from shading and viscosity solutions. In *Proceedings of ICCV*, 2003. 4
- [37] E. Prados and O. Faugeras. Shape from shading: A well-posed problem? In *Proceedings of CVPR*, 2005. 4
- [38] R. Ramamoorthi and P. Hanrahan. An Efficient Representation for Irradiance Environment Maps. In *Proceedings of SIGGRAPH*, 2001. 3
- [39] S. R. Richter and S. Roth. Discriminative shape from shading in uncalibrated illumination. In *Proceedings of CVPR*, 2015. 4
- [40] E. Rouy and A. Tourin. A viscosity solutions approach to shape-from-shading. *SIAM Journal on Numerical Analysis*, 29(3):867–884, 1992. 4
- [41] D. Samaras, D. Metaxas, P. Fua, and Y. G. Leclerc. Variable albedo surface reconstruction from stereo and shape from shading. In *Proceedings of CVPR*, 2000. 2
- [42] A. Tankus, N. A. Sochen, and Y. Yeshurun. A new perspective (on) shape-from-shading. In *Proceedings of ICCV*, 2003. 4
- [43] E. Tola, V. Lepetit, and P. Fua. DAISY: An Efficient Dense Descriptor Applied to Wide Baseline Stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):815–830, 2010. 2
- [44] G. Vogiatzis, P. Torr, and R. Cippola. Multi-view stereo via volumetric graph-cuts. In *Proceedings of CVPR*, 2005. 2
- [45] C. Wu, B. Wilburn, Y. Matsushita, and C. Theobalt. High-quality shape from multi-view stereo and shading under general illumination. In *Proceedings of CVPR*, 2011. 2
- [46] C. Wu, M. Zollhöfer, M. Nießner, M. Stamminger, S. Izadi, and C. Theobalt. Real-time shading-based refinement for consumer depth cameras. *ACM Transactions on Graphics*, 33(6):200:1–200:10, 2014. 5
- [47] R. Zhang, P.-S. Tsai, J. E. Cryer, and M. Shah. Shape-from-shading: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(8):690–706, 1999. 2, 3
- [48] M. Zollhöfer, A. Dai, M. Innman, C. Wu, M. Stamminger, C. Theobalt, and M. Nießner. Shading-based refinement on volumetric signed distance functions. *ACM Transactions on Graphics*, 34(4):96:1–96:14, 2015. 1, 2, 8