Photometric Depth Super-Resolution

Bjoern Haefner*, Songyou Peng*, Alok Verma*, Yvain Quéau, and Daniel Cremers

Abstract—This study explores the use of photometric techniques (shape-from-shading and uncalibrated photometric stereo) for upsampling the low-resolution depth map from an RGB-D sensor to the higher resolution of the companion RGB image. A single-shot variational approach is first put forward, which is effective as long as the target's reflectance is piecewise-constant. It is then shown that this dependency upon a specific reflectance model can be relaxed by focusing on a specific class of objects (e.g., faces), and delegate reflectance estimation to a deep neural network. A multi-shots strategy based on randomly varying lighting conditions is eventually discussed. It requires no training or prior on the reflectance, yet this comes at the price of a dedicated acquisition setup. Both quantitative and qualitative evaluations illustrate the effectiveness of the proposed methods on synthetic and real-world scenarios.

Index Terms—RGB-D cameras, depth super-resolution, shape-from-shading, photometric stereo, variational methods, deep learning.

1 Introduction

RGB-D sensors have become very popular for 3D-reconstruction, in view of their low cost and ease of use. They deliver a colored point cloud in a single shot, but the resulting shape often misses thin geometric structures. This is due to noise, quantisation and, more importantly, the coarse resolution of the depth map. In comparison, the quality and resolution of the companion RGB image are substantially better. For instance, the Asus Xtion Pro Live device delivers 1280×1024 px^2 RGB images, but only up to 640×480 px^2 depth maps. The depth map thus needs to be up-sampled to the same resolution of the RGB image, and the latter could be analysed photometrically to reveal fine-scale details. However, super-resolution of a solitary depth map without additional constraint is an ill-posed problem, and retrieving geometry from either a single color image (shape-from-shading) or from a sequence of color images acquired under unknown, varying lighting (uncalibrated photometric stereo) is another ill-posed problem.

This study explores the resolution of both ill-posedness issues by jointly performing depth super-resolution and photometric 3D-reconstruction. We call this combined approach *photometric depth super-resolution*. It is motivated by the observation that ill-posedness in depth super-resolution and in photometric 3D-reconstruction have different origins. In depth super-resolution, constraints on high-frequency shape variations are missing, while low-frequency (e.g., concave-convex or bas-relief) ambiguities arise in photometric 3D-reconstruction. Therefore, the low-frequency geometric information necessary to disambiguate photometric 3D-reconstruction should be extracted from the low-resolution depth measurements and, symmetrically, the high-resolution photometric clues in the RGB image(s) should provide the high-frequency information required to disambiguate depth super-resolution. One hand thus washes the other: ill-posedness in depth super-resolution is fought using photometric 3D-reconstruction, and vice-versa.

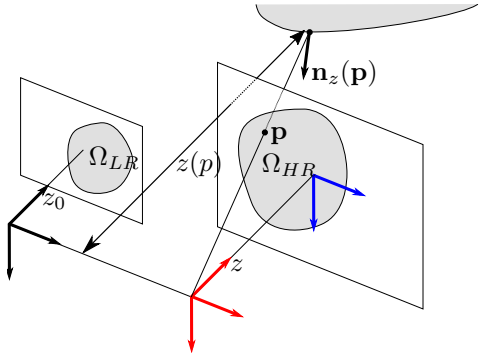


Fig. 1: Schematic representation of the geometric setup. Depth measurements $z^0 : \Omega_{LR} \to \mathbb{R}$ are available over a low-resolution set Ω_{LR}, and color measurements $\mathbf{I} : \Omega_{LR} \to \mathbb{R}^3$ over a high-resolution set Ω_{HR}. Photometric depth super-resolution consists in combining these measurements into a high-resolution depth map $z : \Omega_{HR} \to \mathbb{R}$ over the high-resolution set Ω_{HR}. The red (RGB sensor) frame is the reference coordinates system for 3D-measurements, the blue one is the reference coordinates system for pixel coordinates, and $\mathbf{n}_{z,\nabla z}(\mathbf{p}) \in \mathbb{S}^2 \subset \mathbb{R}^3$ is the unit-length normal to the surface at the 3D-point conjugate to pixel $\mathbf{p} \in \Omega_{HR}$. This normal vector implicitly depends upon the unknown depth $z(\mathbf{p})$ and its gradient $\nabla z(\mathbf{p})$, see Eq. (2).

A generic RGB-D sensor is considered, which consists of a depth sensor and an RGB camera facing a surface, aligned in such a way that both of their optical axes are parallel and both of their optical centers lie on a plane orthogonal to these axes (see Figure 1).

The images of the surface on the focal planes of the depth and the color cameras are denoted respectively by $\Omega_{LR} \subset \mathbb{R}^2$ and $\Omega_{HR} \subset \mathbb{R}^2$. In a single shot, the RGB-D

* *Those authors contributed equally*

- *B. Haefner, A. Verma, and D. Cremers are with the Department of Computer Science, Technical University of Munich, 80333, Germany. E-mail: {bjoern.haefner,alok.verma,cremers}@tum.de*
- *S. Peng is with Advanced Digital Sciences Center, University of Illinois at Urbana-Champaign, Singapore, 138602. E-mail: songyou.peng@adsc-create.edu.sg*
- *Y. Quéau is with the GREYC laboratory, UMR CNRS 6072, Caen, France, and with L@bISEN Yncrea-Ouest, ISEN Brest, France. E-mail: yvain.queau@gmail.com*

Manuscript received Month dd, yyyy; revised Month dd, yyy.

sensor provides two 2D-representations of the surface:

- A geometric one, taking the form of a mapping $z^0 : \Omega_{LR} \to \mathbb{R}$ between pixels in Ω_{LR} and the depth of their conjugate 3D-points on the surface;
- A photometric one, taking the form of a mapping $\mathbf{I} : \Omega_{HR} \to \mathbb{R}^3$ between pixels in Ω_{HR} and the radiance (relatively to the red, green and blue channels of the color camera) of their conjugate 3D-point.

In real-world scenarios, the sets Ω_{LR} and Ω_{HR} are discrete, and the cardinality $|\Omega_{LR}|$ of Ω_{LR} is lower than that $|\Omega_{HR}|$ of Ω_{HR}, since the color camera typically has a higher resolution than the depth sensor. To obtain the richest surface representation, one should thus project the depth measurements z^0 from Ω_{LR} to Ω_{HR}, i.e. estimate a new, high resolution depth map $z : \Omega_{HR} \to \mathbb{R}$. This requires a few assumptions on the relationships between these maps, which are stated in the next paragraphs.

1.1 Assumptions and Problem Statement

Given the assumptions above on the alignment of the depth sensors, and neglecting occlusions, the low-resolution depth map z^0 can be considered as a downsampled version of the sought high-resolution one z, after warping and averaging:

$$z^0 = Kz + \eta_z, \tag{1}$$

with η_z the realisation of a stochastic process representing measurement errors and quantisation, and $K : \mathbb{R}^{|\Omega_{HR}|} \to \mathbb{R}^{|\Omega_{LR}|}$ a linear operator combining warping, blurring and downsampling [1], which can be calibrated beforehand [2]. Solving (1) in terms of the high-resolution depth map z constitutes the *depth super-resolution* problem. Since K is not invertible, further assumptions on the smoothness of the unknown depth map z are required. In this work, we require z to be smooth, in the sense that the normal to the surface exists in every visible point. Denoting by $f > 0$ the focal length of the color camera, and by $\mathbf{p} : \Omega_{HR} \to \mathbb{R}^2$ the field of pixel coordinates with respect to its principal point, the normals to the surface can be parameterised in the image domain as a field $\mathbf{n}_{z,\nabla z} : \Omega_{HR} \to \mathbb{S}^2 \subset \mathbb{R}^3$ of unit-length vectors, which is defined as (see, e.g., [3])

$$\mathbf{n}_{z,\nabla z} = \frac{1}{\sqrt{|f\,\nabla z|^2 + (-z - \mathbf{p}\cdot\nabla z)^2}} \begin{bmatrix} f\,\nabla z \\ -z - \mathbf{p}\cdot\nabla z \end{bmatrix} \tag{2}$$

We further assume that the surface is Lambertian and lit by a collection of infinitely distant point light sources. The irradiance in channel $\star \in \{R, G, B\}$ then writes

$$I_\star = \int_\lambda \int_\omega c_\star(\lambda)\rho(\lambda)\phi(\lambda,\omega)\max\{0, \mathbf{s}(\omega)\cdot\mathbf{n}_{z,\nabla z}\}\,\mathrm{d}\omega\,\mathrm{d}\lambda, \tag{3}$$

where integration is carried out over all wavelengths λ (ρ is the spectral reflectance of the surface and $c_\star$ is the transmission spectrum of the camera in channel $\star$) and all incident lighting directions ω ($\mathbf{s}(\omega)$ is the unit-length vector pointing towards the light source located in direction ω, and $\phi(\cdot,\omega)$ is the spectrum of this source), and where $\mathbf{n}_{z,\nabla z}$ is the unit-length surface normal defined in (2). Assuming achromatic lighting i.e., $\phi(\cdot,\omega) := \phi(\omega)$, using a first-order[1] spherical harmonics approximation [5], [6] of the inner integral, and assuming (3) is satisfied up to additive noise, we obtain

$$\mathbf{I} = \underbrace{\begin{bmatrix} \int_\lambda c_R(\lambda)\rho(\lambda)\mathrm{d}\lambda \\ \int_\lambda c_G(\lambda)\rho(\lambda)\mathrm{d}\lambda \\ \int_\lambda c_B(\lambda)\rho(\lambda)\mathrm{d}\lambda \end{bmatrix}}_{:=\boldsymbol{\rho}} \mathbf{l} \cdot \underbrace{\begin{bmatrix} \mathbf{n}_{z,\nabla z} \\ 1 \end{bmatrix}}_{:=\mathbf{m}_{z,\nabla z}} + \eta_I, \tag{4}$$

with $\eta_I : \Omega_{HR} \to \mathbb{R}^3$ the realisation of a stochastic process standing for noise, quantisation and outliers, $\mathbf{l} \in \mathbb{R}^4$ the achromatic "light vector", $\boldsymbol{\rho} : \Omega_{HR} \to \mathbb{R}^3$ the albedo (Lambertian reflectance) map, relatively to the camera transmission spectra $\{c_\star\}_{\star\in\{R,G,B\}}$, and $\mathbf{m}_{z,\nabla z} : \Omega_{HR} \to \mathbb{R}^4$ a vector field depending upon the surface normals. Solving (4) in terms of the high-resolution depth map z constitutes the *photometric 3D-reconstruction* problem. Remark that both the reflectance $\boldsymbol{\rho}$ and the lighting $\mathbf{l}$ are unknown and represent hidden variables to estimate.

The aim of *photometric depth super-resolution* is then to infer the high-resolution depth map $z : \Omega_{HR} \to \mathbb{R}$ out of the low-resolution one $z^0 : \Omega_{LR} \to \mathbb{R}$ and of the high-resolution photometric measurements $\mathbf{I} : \Omega_{HR} \to \mathbb{R}^3$, while ensuring consistency with the super-resolution constraint in (1) and with the photometric one in (4).

1.2 Contributions and Paper Organisation

As being discussed in Section 2, the literature is remarkably dense in both the fields of depth super-resolution and photometric 3D-reconstruction. However, their joint solving has not been explored so far, apart from our previous conference papers [7] and [8]. The present study subsumes both these contributions into a thorough discussion on photometric depth super-resolution. Besides extending the model-based approach from [7] and the data-based one from [8] with new experiments and discussions, a novel intermediate approach will be presented, which combines the benefits of model- and data-based strategies.

Our conference paper [7] has highlighted that the key issue in photometric depth super-resolution lies within the estimation of the target's reflectance. Using a single RGB-D pair, it can be carried out in the same time as photometric depth super-resolution using a variational approach to shape-from-shading, as discussed in Section 3. This approach however relies on a strong assumption on the piecewise-constantness of the reflectance map.

It is then shown in Section 4 how to relax this assumption by delegating reflectance estimation to a deep neural network trained offline using photometric stereo. The learnt reflectance can be plugged into the former variational framework, which considerably simplifies the numerics and bypasses the need for an explicit reflectance model.

Still, this approach remains effective only as long as the reflectance of the target is similar to those of the training objects. To get rid of any implicit or explicit reflectance model, a sequence of RGB-D data can be captured under varying, unknown lighting [8]. We discuss this uncalibrated photometric stereo-based strategy in Section 5, before drawing our conclusions and suggesting future research directions in Section 6.

1. This study is straightforward to extend to second-order spherical harmonics. However we did not observe substantial improvement with this extension, hence we discuss only the first-order case, which already captures more than 85% of natural illumination [4].

2 Related Works

Single depth image super-resolution requires solving Equation (1) in terms of the high-resolution depth map z. Since K is not invertible, this is an ill-posed problem: there exist infinitely many choices for interpolating between observations (see Figure 1 in the supplementary material). Disambiguation can be carried out by adding observations obtained from different viewing angles [9], [10], [11]. In the more challenging case of a single viewing angle, a smoothness prior on the high-resolution depth map can be added and a variational approach can be followed [1]. One may also resort to machine learning techniques relying on a dictionary of low- and high-resolution depth or edge patches [12], [13]. Such a dictionary can even be constructed from a single depth image by looking for self-similarities [14], [15]. Nevertheless, learning-based depth super-resolution methods remain prone to over-fitting [16], which can also be avoided by combining the respective merits of machine learning and variational approaches [17], [18].

Shape-from-shading [19], [20], [21], [22] is another classical inverse problem which aims at inferring shape from a single image of a scene, by inverting an image formation model such as (4). Common numerical strategies for this task include variational [23], [24] and PDE methods [25], [26], [27], [28]. However, even when reflectance and lighting are known, shape-from-shading is ill-posed. For instance, if assuming for simplicity a graylevel image, frontal lighting, uniform Lambertian reflectance and orthographic projection, shape-from-shading comes down to solving the eikonal equation $|\nabla z| = \sqrt{\frac{1}{I^2} - 1}$ [29]. This equation only provides the magnitude of the depth gradient, and not its direction. The local shape is thus unambiguous in singular points (the tangent vectors in Figure 2 in the supplementary material), but two singular points may either be connected by "going up" or by "going down" (concave / convex ambiguity). Obviously, even more ambiguities arise under more realistic lighting and reflectance assumptions. This is nicely visualised in the "workshop metaphor" of Adelson and Pentland [30]: any image can be explained by a flat shape illuminated uniformly but painted in a complex manner, by a white and frontally-lit surface with a complex geometry, or by a white planar surface illuminated in a complex manner. Shape-from-shading under uniform reflectance but natural lighting has been studied [31], [32], [33], [34], but the case with unknown reflectance requires the introduction of additional priors [35]. This can be avoided by actively controlling the lighting, a variant of shape-from-shading known as photometric stereo which allows to unambiguously estimate shape and reflectance [36]. The problem with uncalibrated lighting is however ill-posed: it can be solved only up to a linear ambiguity [37] which, assuming integrability of the normals, reduces to a generalised bas-relief (GBR) one under directional lighting [38], and to a Lorentz one under spherical harmonics lighting [39]. Resolution of such ambiguities by resorting to additional priors [40], [41], [42] , and extensions to non-Lambertian reflectances [43], remain active research topics. It has also been shown recently in [44] that PDE-based approaches may be worthwhile for uncalibrated photometric stereo, because they implicitly enforce integrability and thus naturally reduce ambiguities.

Shape-from-shading has recently gained new life with the emergence of RGB-D sensors. Indeed, the rough depth map can be used as prior to "guide" shape-from-shading and thus circumvent its ambiguities. This has been achieved in both the multi-view [45], [46], [47] and the single-shot [48], [49], [50], [51], [52], [53] cases. Still, the resolutions of the input image and depth map are assumed equal, and the same holds for approaches resorting to photometric stereo instead of shape-from-shading [54], [55], [56], [57]. In fact, depth super-resolution and photometric 3D-reconstruction have been widely studied, but rarely together. Several methods were proposed to coalign the depth edges in the super-resolved map with edges in the high-resolution color image [2], [58], [59], [60], [61], but such approaches only consider sparse color features and may thus miss thin geometric structures. Some authors super-resolve the photometric stereo results [62], and others generate high-resolution images using photometric stereo [63], but none employ low-resolution depth clues except those of [64], who combine calibrated photometric stereo with structured light sensing. However, this involves a non-standard setup and careful lighting calibration, and reflectance is assumed to be uniform. Such issues are circumvented in the building blocks [7] and [8] of this study, which deal with photometric depth super-resolution based on, respectively, shape-from-shading and photometric stereo. Let us start by presenting the former approach, which is a single-shot solution to photometric depth super-resolution based on a variational approach to shape-from-shading.

3 Single Shot Depth Super-Resolution using Shape-from-Shading

In this section, we assume that the only available data is the input image $\mathbf{I}$ as well as the low-resolution depth map z^0, and we aim at jointly solving the ill-posed problems of depth super-resolution and shape-from-shading. To obtain a high-resolution depth map z as described in (1) as well as encoded in the normals in (4), we opt for a Bayesian-to-variational strategy [65].

3.1 Bayesian-to-Variational Rationale

Besides the high-resolution depth map z, neither the reflectance $\boldsymbol{\rho}$ nor the lighting vector $\mathbf{l}$ is known. We treat the joint recovery of these three quantities as a maximum a posteriori (MAP) estimation problem. To this end we aim at maximising the posterior distribution of $\mathbf{I}$ and z^0 which, according to Bayes rule, writes

$$\mathcal{P}(z, \boldsymbol{\rho}, \mathbf{l} | z^0, \mathbf{I}) = \frac{\mathcal{P}(z^0, \mathbf{I} | z, \boldsymbol{\rho}, \mathbf{l}) \mathcal{P}(z, \boldsymbol{\rho}, \mathbf{l})}{\mathcal{P}(z^0, \mathbf{I})}. \tag{5}$$

In (5), the denominator is the marginal likelihood, which is a constant wrt. the variables $z, \boldsymbol{\rho}$ and $\mathbf{l}$ and can thus be neglected during optimisation. The numerator is the product of the likelihood $\mathcal{P}(z^0, \mathbf{I} | z, \boldsymbol{\rho}, \mathbf{l})$ and the prior distribution $\mathcal{P}(z, \boldsymbol{\rho}, \mathbf{l})$, which both need to be further discussed.

The measurements of depth and image observations being done using separate sensors, z^0 and $\mathbf{I}$ are statistically independent and thus the likelihood factors out as $\mathcal{P}(z^0, \mathbf{I} | z, \boldsymbol{\rho}, \mathbf{l}) = \mathcal{P}(z^0 | z, \boldsymbol{\rho}, \mathbf{l}) \, \mathcal{P}(\mathbf{I} | z, \rho, \mathbf{l})$. Furthermore, we

assume that the process of how the depth map z^0 is acquired is depending neither on lighting $\mathbf{l}$ nor on reflectance $\boldsymbol{\rho}$. Given this, the marginal likelihood for the depth map z^0 can be written as $\mathcal{P}(z^0|z,\boldsymbol{\rho},\mathbf{l}) = \mathcal{P}(z^0|z)$. Assuming that noise η_z in Eq. (1) is homoskedastic, zero-mean and Gaussian-distributed with variance σ_z^2, we further have $\mathcal{P}(z^0|z) \propto \exp\left\{-\frac{\|Kz-z^0\|^2_{\ell^2(\Omega_{LR})}}{2\sigma_z^2}\right\}$. Concerning the marginal likelihood of $\mathbf{I}$, we assume the random variable η_I in (4) follows a homoskedastic Gaussian distribution with zero mean and covariance matrix $\mathrm{diag}(\sigma_I^2,\sigma_I^2,\sigma_I^2) \in \mathbb{R}^{3\times 3}$, and thus $\mathcal{P}(\mathbf{I}|z,\boldsymbol{\rho},\mathbf{l}) \propto \exp\left\{-\frac{\|\boldsymbol{\rho}\langle \mathbf{m}_{z,\nabla z},\mathbf{l}\rangle - \mathbf{I}\|^2_{\ell^2(\Omega_{HR})}}{2\sigma_I^2}\right\}$, where $\langle\cdot,\cdot\rangle$ is the standard scalar product in $\mathbb{R}^4$. Given the above derivation, the likelihood in (5) is given by

$$\mathcal{P}(z^0,\mathbf{I}|z,\boldsymbol{\rho},\mathbf{l}) \propto \exp\Big\{-\frac{\|Kz-z^0\|^2_{\ell^2(\Omega_{LR})}}{2\sigma_z^2} - \frac{\|\boldsymbol{\rho}\langle \mathbf{m}_{z,\nabla z},\mathbf{l}\rangle - \mathbf{I}\|^2_{\ell^2(\Omega_{HR})}}{2\sigma_I^2}\Big\} \tag{6}$$

The prior distribution $\mathcal{P}(z,\boldsymbol{\rho},\mathbf{l})$ in (5) can be derived in a similar manner. We assume that geometry (z), reflectance ($\boldsymbol{\rho}$) and lighting ($\mathbf{l}$) are statistically independent[2], thus the prior distribution factors out as

$$\mathcal{P}(z,\boldsymbol{\rho},\mathbf{l}) = \mathcal{P}(z)\mathcal{P}(\boldsymbol{\rho})\mathcal{P}(\mathbf{l}). \tag{7}$$

Lighting is modeled as a low-frequency quantity over the whole image domain, 4-dimensional (1st order harmonics) in our case, cf. (4). This allows us to use an improper prior for the lighting:

$$\mathcal{P}(\mathbf{l}) = \text{constant}. \tag{8}$$

The prior on z is slightly more evolved. As we want to prevent oversmoothing (ℓ^2 regularisation) and/or staircasing artefacts (ℓ^1 regularisation), we make use of a minimal surface prior [66]. To this end, a parametrisation $\mathrm{d}\mathcal{A}_{z,\nabla z} : \Omega_{HR} \to \mathbb{R}$ mapping each pixel to the corresponding area of the surface element is required. This writes $\mathrm{d}\mathcal{A}_{z,\nabla z} = \frac{z}{f^2}\sqrt{|f\,\nabla z|^2 + (-z-\mathbf{p}\cdot\nabla z)^2}$, and the total surface area is then given by $\|\mathrm{d}\mathcal{A}_z dz\|_{\ell^1(\Omega_{HR})}$. Introducing a free parameter $\alpha > 0$ to control the surface smoothness, the minimal surface prior can then be stated as

$$\mathcal{P}(z) \propto \exp\left\{-\frac{\|\mathrm{d}\mathcal{A}_{z,\nabla z}\|_{\ell^1(\Omega_{HR})}}{\alpha}\right\}. \tag{9}$$

Following the Retinex theory [67], reflectance $\boldsymbol{\rho}$ can be assumed piecewise-constant, resulting in a Potts prior

$$\mathcal{P}(\boldsymbol{\rho}) \propto \exp\left\{-\frac{\|\nabla\boldsymbol{\rho}\|_{\ell^0(\Omega_{HR})}}{\beta}\right\}, \tag{10}$$

with $\beta > 0$ controlling the degree of discontinuities in the reflectance $\boldsymbol{\rho}$. Note that $\boldsymbol{\rho}$ is a vector field, thus for each pixel $\mathbf{p}$, $\nabla\boldsymbol{\rho}(\mathbf{p}) = [\nabla\rho_R(\mathbf{p}),\nabla\rho_G(\mathbf{p}),\nabla\rho_B(\mathbf{p})]^\top \in \mathbb{R}^{3\times 2}$, and we use the following definition of the $\|\cdot\|_{\ell^0(\Omega_{HR})}$ "norm": $\|\boldsymbol{\rho}\|_{\ell^0(\Omega_{HR})} := \sum_{\mathbf{p}\in\Omega_{HR}} \begin{cases} 0 & \text{if } |\boldsymbol{\rho}(\mathbf{p})|_F = 0, \\ 1 & \text{else} \end{cases}$, with $|\cdot|_F$ the Frobenius norm over $\mathbb{R}^{3\times 2}$.

The MAP estimate $(z^*,\boldsymbol{\rho}^*,\mathbf{l}^*)$ for depth, reflectance and lighting is eventually attained by maximising the posterior distribution (5) or, equivalently, minimising its negative logarithm. Plugging Eqs. (6) to (10) into (5), and discarding all (additive) constants, the joint estimation of depth, reflectance and lighting comes down to solving the following variational problem:

$$\begin{matrix}\min\limits_{\substack{z:\Omega_{HR}\to\mathbb{R}\\ \boldsymbol{\rho}:\Omega_{HR}\to\mathbb{R}^3\\ \mathbf{l}\in\mathbb{R}^4}}\end{matrix} \|\boldsymbol{\rho}\,\langle \mathbf{m}_{z,\nabla z},\mathbf{l}\rangle - \mathbf{I}\|^2_{\ell^2(\Omega_{HR})} + \mu\,\|Kz-z^0\|^2_{\ell^2(\Omega_{LR})} + \nu\,\|\mathrm{d}\mathcal{A}_{z,\nabla z}\|_{\ell^1(\Omega_{HR})} + \lambda\,\|\nabla\boldsymbol{\rho}\|_{\ell^0(\Omega_{HR})}, \tag{11}$$

where the trade-off parameters (μ,ν,λ) are given by

$$\mu = \frac{\sigma_I^2}{\sigma_z^2}, \qquad \nu = \frac{\sigma_I^2}{\alpha}, \qquad \lambda = \frac{\sigma_I^2}{\beta}. \tag{12}$$

3.2 Numerical Solving of (11)

The variational problem in (11) is not only nonconvex, but also inherits a nonlinear dependency upon the gradient of z, see (4) along with (2). Compared to other methods, which overcome this issue by either following a two-step approach via optimising over the normals and then fitting an integrable surface to it [48] (a strategy which may fail if the estimated normals are non-integrable), or by freezing the nonlinearity [51] (which may yield convergence issues, in view of the nonconvexity of the optimisation problem), we solve for the depth directly and without any approximation. To this end we introduce an auxiliary vector field $\boldsymbol{\theta} : \Omega_{HR} \to \mathbb{R}^3$ with $\boldsymbol{\theta} := (z,\nabla z)$, which separates the nonlinearities induced by the shape-from-shading and minimal surface terms from the dependency upon the gradient induced by the normal calculation, see [34]. This splitting approach can formally be written as a constrained optimisation problem of the form

$$\begin{aligned}\begin{matrix}\min\limits_{\substack{z:\Omega_{HR}\to\mathbb{R}\\ \boldsymbol{\rho}:\Omega_{HR}\to\mathbb{R}^3\\ \mathbf{l}\in\mathbb{R}^4\\ \boldsymbol{\theta}:\Omega_{HR}\to\mathbb{R}^3}}\end{matrix} & \|\boldsymbol{\rho}\,\langle \mathbf{m}_{\boldsymbol{\theta}},\mathbf{l}\rangle - \mathbf{I}\|^2_{\ell^2(\Omega_{HR})} + \mu\,\|Kz-z^0\|^2_{\ell^2(\Omega_{LR})} \\ & + \nu\,\|\mathrm{d}\mathcal{A}_{\boldsymbol{\theta}}\|_{\ell^1(\Omega_{HR})} + \lambda\,\|\nabla\boldsymbol{\rho}\|_{\ell^0(\Omega_{HR})} \\ & \text{s.t. } \boldsymbol{\theta} = (z,\nabla z).\end{aligned} \tag{13}$$

To solve the nonconvex, nonsmooth and constrained optimisation problem (13) we make use of a multi-block ADMM scheme [68], [69], [70]. This comes down to iterating a sequence consisting of minimisations of the augmented Lagrangian

$$\begin{aligned}\mathcal{L}(z,\boldsymbol{\rho},\mathbf{l},\boldsymbol{\theta},\mathbf{u}) = {} & \|\boldsymbol{\rho}\,\langle \mathbf{m}_{\boldsymbol{\theta}},\mathbf{l}\rangle - \mathbf{I}\|^2_{\ell^2(\Omega_{HR})} + \mu\,\|Kz-z^0\|^2_{\ell^2(\Omega_{LR})} \\ & + \nu\,\|\mathrm{d}\mathcal{A}_{\boldsymbol{\theta}}\|_{\ell^1(\Omega_{HR})} + \lambda\,\|\nabla\boldsymbol{\rho}\|_{\ell^0(\Omega_{HR})} \\ & + \langle \mathbf{u},\boldsymbol{\theta}-(z,\nabla z)\rangle + \frac{\kappa}{2}\,\|\boldsymbol{\theta}-(z,\nabla z)\|^2_{\ell^2(\Omega_{HR})}\end{aligned} \tag{14}$$

over the primal variables z, $\boldsymbol{\rho}$, $\mathbf{l}$ and $\boldsymbol{\theta}$, and one gradient ascent step over the dual variable $\mathbf{u} : \Omega_{HR} \to \mathbb{R}^3$ ($\kappa > 0$ can be viewed as a step size).

2. The non-dependency of reflectance upon geometry and lighting follows immediately from the Lambertian assumption, as the surface is reflecting light equally in all directions. Independence of geometry and lighting follows from the distant-light assumption and neglection of self-reflections.

At iteration (k), one sweep of this scheme writes as:

$$\boldsymbol{\rho}^{(k+1)} = \underset{\boldsymbol{\rho}:\Omega_{HR}\to\mathbb{R}^3}{\arg\min} \left\| \boldsymbol{\rho} \left\langle \mathbf{m}_{\boldsymbol{\theta}^{(k)}}, \mathbf{l}^{(k)} \right\rangle - \mathbf{I} \right\|^2_{\ell^2(\Omega_{HR})} + \lambda \left\| \nabla \boldsymbol{\rho} \right\|_{\ell^0(\Omega_{HR})}, \tag{15}$$

$$\mathbf{l}^{(k+1)} = \underset{\mathbf{l}\in\mathbb{R}^4}{\arg\min} \left\| \boldsymbol{\rho}^{(k+1)} \left\langle \mathbf{m}_{\boldsymbol{\theta}^{(k)}}, \mathbf{l} \right\rangle - \mathbf{I} \right\|^2_{\ell^2(\Omega_{HR})}, \tag{16}$$

$$\boldsymbol{\theta}^{(k+1)} = \underset{\boldsymbol{\theta}:\Omega_{HR}\to\mathbb{R}^3}{\arg\min} \left\| \boldsymbol{\rho}^{(k+1)} \left\langle \mathbf{m}_{\boldsymbol{\theta}}, \mathbf{l}^{(k+1)} \right\rangle - \mathbf{I} \right\|^2_{\ell^2(\Omega_{HR})} + \nu \left\| \mathrm{d}\mathcal{A}_{\boldsymbol{\theta}} \right\|_{\ell^1(\Omega_{HR})} + \frac{\kappa}{2} \left\| \boldsymbol{\theta} - (z, \nabla z)^{(k)} + \mathbf{u}^{(k)} \right\|^2_{\ell^2(\Omega_{HR})}, \tag{17}$$

$$z^{(k+1)} = \underset{z:\Omega_{HR}\to\mathbb{R}}{\arg\min}\ \mu \left\| Kz - z^0 \right\|^2_{\ell^2(\Omega_{LR})} + \frac{\kappa}{2} \left\| \boldsymbol{\theta}^{(k+1)} - (z, \nabla z) + \mathbf{u}^{(k)} \right\|^2_{\ell^2(\Omega_{HR})}, \tag{18}$$

$$\mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} + \boldsymbol{\theta}^{(k+1)} - (z, \nabla z)^{(k+1)}. \tag{19}$$

The albedo subproblem (15) is solved using the primal-dual algorithm [71]. The lighting update step in (16) is done using the pseudo-inverse. The $\boldsymbol{\theta}$-update (17) is a nonlinear optimisation subproblem, yet free of neighboring pixel dependency thanks to the proposed splitting. It can be solved independently in each pixel using the implementation [72] of the L-BFGS method [73]. Eventually, the conjugate gradient method is applied on the normal equations of (18), which is a sparse linear least squares problem.

Our initial values for $(k) = (0)$ are chosen to be $\boldsymbol{\rho}^{(0)} = \mathbf{I}$, $\mathbf{l}^{(0)} = [0, 0, -1, 0]^\top$, $z^{(0)}$ a smoothed version of z^0 using the guided filter [74] followed by bicubic interpolation to upsample to the image domain Ω_{HR}, $\boldsymbol{\theta}^{(0)} = (z, \nabla z)^{(0)}$, $\mathbf{u}^{(0)} = 0$ and $\kappa = 10^{-4}$. Due to the problem being nonsmooth and nonconvex, to date no convergence results have been provided and we leave this as future work. Nevertheless, in our experiments we have never encountered any problem reaching convergence, which we consider as reached if the relative residual falls below some threshold:

$$r_{\text{rel}} := \frac{\left\| z^{(k+1)} - z^{(k)} \right\|_{\ell^2(\Omega_{HR})}}{\left\| z^{(0)} \right\|_{\ell^2(\Omega_{HR})}} < 10^{-5}, \tag{20}$$

and if the constraint $\boldsymbol{\theta} = (z, \nabla z)$ is numerically satisfied, i.e.

$$r_{\text{c}} := \left\langle \mathbf{u}^{(k+1)}, \boldsymbol{\theta}^{(k+1)} - (z, \nabla z)^{(k+1)} \right\rangle + \frac{\kappa}{2} \left\| \boldsymbol{\theta}^{(k+1)} - (z, \nabla z)^{(k+1)} \right\|^2_{\ell^2(\Omega_{HR})} < 5 \cdot 10^{-6}. \tag{21}$$

To ensure the latter, the step size κ is multiplied by a factor of 1.5 after each iteration.

The scheme is implemented in Matlab, except the albedo update (15) which is implemented in CUDA. Depending on the datasets, convergence is reached between $10s$ and $90s$.

3.3 Experiments

We used publicly available datasets (Figure 3 in the supplementary material) in order to determine appropriate values for the hyper-parameters (μ, ν, λ) involved in Problem (11), and for comparing against the state-of-the-art. To this end, 3D-meshes were rendered into high-resolution ground-truth depth maps of size $[480 \times 640\ \text{px}^2]$, then into low-resolution depth maps with scale factors of 8, 4 and 2. Additive zero-mean Gaussian noise with standard deviation 10^{-3} times the squared original depth value (consistently with real-world measurements from [75]) is then added to the low-resolution depth maps, before quantisation. High-resolution RGB images were rendered from the ground-truth depth map using the first-order spherical harmonics model (4), with $\mathbf{l} = [0, 0, -1, 0.2]^\top$, four different high-resolution reflectance maps and an additive zero-mean Gaussian noise with standard deviation 1% the maximum intensity. For quantitative evaluation, we consider the root mean squared error (RMSE) on the estimated depth and reflectance maps, and the mean angular error (MAE) on surface normals.

Although the optimal value of each parameter can be deduced using (12), it can be difficult to estimate the statistics in practice, thus we consider (μ, ν, λ) as tunable hyperparameters. To select an appropriate set of values for them, we initially set $\mu = 0.5$, $\nu = 0.01$ and $\lambda = 1$. We then evaluate the impact of each parameter by varying it while keeping the remaining two fixed. The results of this experiment (see Figure 4 in the supplementary material) are as follows. Large values of μ force the depth map to keep close to the noisy input, while small values make the depth prior less important so not capable of disambiguating shape-from-shading. Inbetween, the range $\mu \in [10^{-1}, 10]$ seems to provide appropriate results. As for ν, large values produce over-smoothed results and small ones result in slightly noisier depth estimates, although the albedo estimate seems unaffected by this choice. Overall, the range $\nu \in [1, 10^2]$ seems appropriate, although depth regularisation really matters only if color cannot be exploited, e.g. due to shadows, black reflectance, saturation or strong noise in the RGB image (see real-world experiments). The parameter λ strongly impacts both the resulting albedo and depth: too small (resp., high) values for λ result in over (resp., under)-segmentation problems, and in both cases shading information gets propagated to the albedo. We found $\lambda \in [10^{-1}, 10]$ to be a reasonable choice. Overall, we opted for the set of parameters $(\mu, \nu, \lambda) = (0.1, 0.7, 1)$ for our experiments.

Using this set of parameters, we then conducted qualitative and quantitative comparison of our method against two other single-shot approaches, i.e. a learning-based approach [13][3] and an image-based approach [60]. To emphasise the interest of joint shape-from-shading and depth super-resolution over shading-based depth refinement using downsampled images, we also consider [51]. Table 1 and Figure 5, both in the supplementary material, show comparisons of our results with those of these methods. As can be clearly seen, our method systematically overcomes the competitors in terms of MAE values, which indicates that fine-scale (high-frequency) geometric details are better recovered. The RMSE on depth rather evaluates the overall (low-frequency) adequation with ground-truth, and for this metric our results are comparable with [13], which achieves the best results. Overall, the proposed method provides the best compromise between the recovery of high- and low-frequency geometric information.

3. [13] is learning-based and needs to be trained for a certain upsampling factor. Since the authors only provide trained data for a factor of 4, this method was evaluated only for this factor.

Real-world experiments were then carried on publicly available datasets [46], [47], [76], [77], as well as data we captured ourselves with an Intel RealSense D415 and an Asus Xtion Pro Live camera. We acquired RGB images of size $[1280 \times 720\ \text{px}^2]$ and $[1280 \times 1024\ \text{px}^2]$, respectively, and depth images of resolution $[640 \times 480\ \text{px}^2]$ and $[320 \times 240\ \text{px}^2]$, both at 30fps. Data was captured indoor with an LED attached to the camera in order to reinforce shading in the RGB images. The objects of interest were manually segmented from background before processing.

Fig. 2: Results on real-world datasets "Android", "Basecap", "Minion" and "Rucksack" we captured with an Intel Realsense D415 camera. All datasets shown here have a scaling factor of 2 between the low-resolution depth z^0 and the high-resolution image $\mathbf{I}$.

Figure 2 shows the resulting estimates of ρ and z on real-world data. More qualitative results can be found in the supplementary material. The "Android" dataset shows that our approach nicely separates albedo and depth estimation. All shading information is explained with geometry as our Potts prior prevents shading information being propagated into reflectance. Whenever color gets saturated or too low (cf. "Basecap" dataset), the minimal surface drives super-resolution, which adds robustness. Additionally fine details, such as the stitches on the peak or the rivet of the bottle opener of the cap are nicely recovered. The geometry of the 3-dimensional stitching "GUINNES" is not explained as albedo information, thus it is correclty explained in z. Over-segmentation of reflectance can be seen in the "minion" dataset (the eyes, the "Gru" logo central of the dungarees and the left foot). Thanks to the minimal surface prior the depth estimate does not seem to deteriorate greatly. For the "Rucksack" dataset even fine details, such as the wrinkles can be recovered and although the material is not purely lambertian, our algorithm seems to cope well with it.

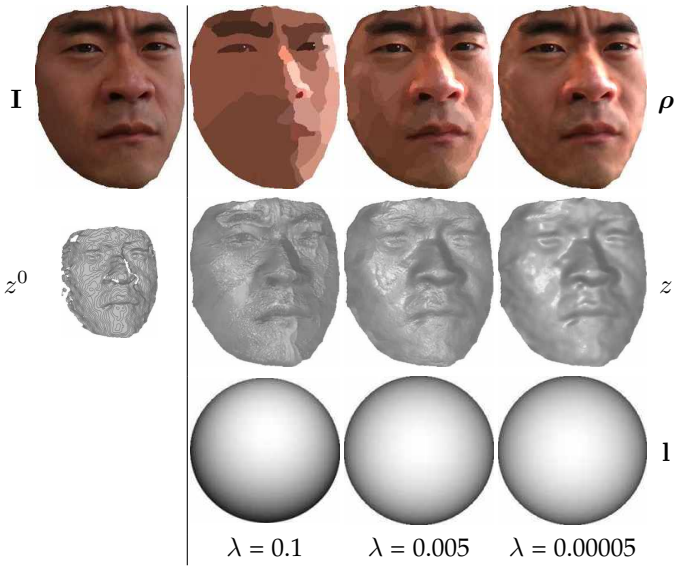


Fig. 3: Failure case of the single-shot approach. Left: input data, for a human face whose reflectance does not match our piecewise-constant prior. Right: estimated reflectance, depth and lighting, for different values of the parameter λ, which controls the Potts regulariser. The Potts prior yields artefacts in the depth map, yet relaxing it prevents one from refining geometry using shape-from-shading.

Our method only fails when reflectance does not fit the Potts prior, e.g. on faces. For such objects with smoothly varying reflectance the piecewise-constant albedo assumption induces bias in the estimated depth. Indeed, the prior forbids to explain thin brightness variations in terms of reflectance, and thus the depth is forced to account for them. This results in noisy high-resolution depth maps (second row in Figure 3). Of course, one could relax the piecewise-constant reflectance assumption by appropriately tuning the weight of the Potts prior, i.e. by reducing the value of λ. However, this causes the albedo map to entirely explain the input image. Consequently, shading is assumed to be negligible, and there is no hope to properly refine the input depth map through shape-from-shading, cf. columns three and four in Figure 3.

This failure case illustrates the difficulty of designing a Bayesian prior which would properly split geometry and albedo information. The rest of this manuscript discusses two different strategies to circumvent this issue: by replacing the albedo estimation brick of the proposed variational framework with a deep neural network, or by acquiring additional data. The former approach is described in the next section.

4 Depth Super-Resolution using Shape-from-Shading and Reflectance Learning

The need for a strong prior on the target's reflectance is a serious bottleneck in single-shot depth super-resolution using shape-from-shading. To circumvent this issue, we investigate in this section the combination of a deep learning strategy (to estimate reflectance) with a simplified version of the proposed variational framework (to carry out depth super-resolution, with known reflectance).

4.1 Motivations and Construction of our Method

If we replace the assumption of piecewise-constant albedo by the much stronger assumption of known albedo, the variational problem from the previous section comes down to jointly achieving depth super-resolution and low-order lighting estimation, and is thus simplified substantially. Yet, the task of designing a reflectance prior which is both realistic and numerically tractable is replaced with that of designing an efficient method for estimating a reflectance map out of a high-resolution RGB image. Luckily, this problem has long been investigated in the computer vision community: it is an intrinsic image decomposition problem. Some variational solutions exist [35], [78], yet they rely on explicit reflectance priors and thus suffer from the same limitations as the previously proposed approach. One recent alternative is to rather resort to convolutional neural networks (CNNs), see for instance [79].

However, one important issue pertaining to learning-based approaches is the lack of inter-class generalisation (see Figure 9 in the supplementary material). Nevertheless, as long as the object to analyse resembles those used during the training stage, the albedo estimates are satisfactory. Therefore, our proposal is to replace our man-made reflectance prior (piecewise constantness) by a less explicit prior on the class of objects the target belongs to. In this section, we focus on the human faces class, as e.g. in [80], in view both of the richness of geometric details to recover and of the complexity of the reflectance.

Let us emphasise that we resort to CNNs only for reflectance estimation and not for geometry refinement, although several deep learning strategies are able to provide shape clues along with reflectance information [81], [82], [83], [84]. Indeed, such methods have shown commendable results yet they are fraught with good-to-the-eye but possibly physically-incorrect geometry estimates, probably because during testing time they are unfettered by any concrete physics-based model and prior. See, for instance, Figure 10 in the supplementary material. Given that we do already have a physics-based depth refinement framework at hand, which furthermore makes use of the available low-resolution geometric clues from the depth sensor, we believe it is more sound to pick the best from both worlds - deep learning and variational methods. The solution we advocate thus contains two building blocks: a deep neural network prior-lessly learns the mapping from the input RGB image to reflectance for a particular class of objects (here, human faces), and then our variational framework based on shape-from-shading provides a physically-sound numerical framework for depth super-resolution.

4.2 Reflectance Learning

To train a CNN for the estimation of the human face's reflectance, one needs at his disposal hundreds of facial images in vivid lighting and viewing conditions, along with the corresponding albedo maps. This could be achieved using, e.g., photometric stereo, yet the process would be very tedious. Training a neural network using synthetic images is a common and much simpler alternative: for instance, the approach from [85] resorts to the ShapeNet 3D-model library for estimating the albedo of inanimate objects. We follow a similar approach, but dedicated to human faces.

We consider for this purpose the ICT-3DRFE database [76], [77], which comprise of 3D meshes of human faces, reflectance maps and normal maps. These databases were captured using a Light Stage, which provides fine-detailed shape and reflectance. Using a rendering software like Blender, one can then relight the faces and change viewing angles in order to obtain hundreds of shaded RGB images along with ground-truth albedo maps. Our training dataset consists of 21 faces, each enacting 15 different expressions. For each face and each expression, several images are acquired under varying lighting conditions induced by combining ten extended light sources (see Figure 11 in supplementary material). In practice, eight different lighting conditions are simulated by modulating the intensity of each light source, in accordance to the usual lighting in homes and offices, e.g. light sources on the ceiling, walls, windows etc. Furthermore, rendering of the faces is done from three different viewing angles, i.e. center, slight left and slight right. Eventually, the images are generated using both a diffuse and a specular reflectance model. These choices are expected to make the CNN robust to viewpoint and lighting variations. In total, after pruning the dataset and augmenting the faces for lighting, viewpoint and specularity, the training set comprises of 5175 images. Figure 4 shows some rendering examples, along with the corresponding ground-truth albedo maps.

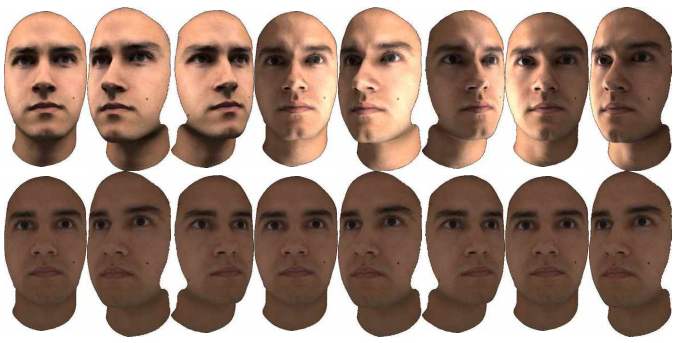


Fig. 4: Examples of human faces rendered under varying viewing and lighting conditions (top), along with the corresponding albedo maps (bottom).

A CNN is then trained to learn the mapping from the rendered face images to the corresponding ground-truth reflectance. Our network architecture is based on U-Net [86]. Generally, U-Net comprises of convolution and nonlinear layers which downsample the input to a 1D array and then upsample to the same input size using transpose convolution and nonlinear layers. Apart from these

layers, an important architectural nuance of U-Net is the skip connections between downsampling and upsampling layers. This allows U-Net to produce sharp results, which is crucial for albedo estimation. For a detailed description of the U-Net architecture used here we refer readers to the supplementary material. Let us emphasise that the architecture of this network is remarkably simple. Once reflectance estimation is dropped out, the variational problem (11) for joint depth super-resolution and lighting estimation also becomes rather simple. Still, the appropriate combination of such simple frameworks does provide state-of-the-art results, as we shall see in the following.

4.3 Experiments

For quantitative evaluation we used two subjects with a total of 30 faces, from the ICT-3DRFE Database [77]. These faces, the applied lighting and camera viewpoint were not used during training. High resolution RGB images were rendered at $[512 \times 512\ \mathrm{px}^2]$ resolution from the ground truth albedo and ground truth depth under first order spherical harmonics lighting $\mathbf{l} = [0, 0, -1, 0.2]^\top$. The low-resolution depth maps were achieved by downsampling the ground truth depth by a factor of 2 and 4. Zero-mean Gaussian noise of 1% was added to the RGB images and zero-mean Gaussian noise with standard deviation of 3.10^{-3} times the squared original depth value is added to the low-resolution depth maps, before quantisation.

Thorough parameter evaluation in this section is not carried and this can be nicely explained with (10) and (12): The scaling parameter β for the prior described in (10) does not affect the parameters ν and μ, thus dropping the reflectance prior does not change the results of the parameter evaluation described in Section 3.3.

We compared the results with two other state-of-the-art methods: SIRFS [35] and SfSNet [84] (both do not perform super-resolution). The former is a prior based Shape-from-Shading optimisation approach which can estimate depth, produce albedo and lighting. The latter is a complete deep learning based approach which only relies on an RGB image and results in an $[128 \times 128\ \mathrm{px}^2]$ normal map, albedo and shading. The quantitative results (Figure 13 and Table 2 in the supplementary material) depict that we systematically outperform other state-of-the-art methods in terms of the MAE value and for larger scaling factors than 2 also for the RMSE. For a scaling factor of 2 our method is on-par with [35] and thus results in the best trade-off between good MAE and RMSE values.

For real-world experiments we captured faces under natural lighting using the Intel RealSense D415 sensor. The RGB image resolution we use is $[1280 \times 720\ \mathrm{px}^2]$ and the depth resolution is $[640 \times 360\ \mathrm{px}^2]$, resulting in a scaling factor of 2. The results are illustrated in Figure 5. They demonstrate the ability to produce detail-preserving depth maps with subtle wrinkles out of a low-quality input depth. Thanks to our variational model, even geometric details of the teeth can be captured in geometry, where most other pure deep learning approaches fail or do not model this behavior at all. Thanks to the deep net albedo prior we are able to handle complex albedo of faces and thus result in high detail geometric estimates, capturing wrinkles and thin facial structures.

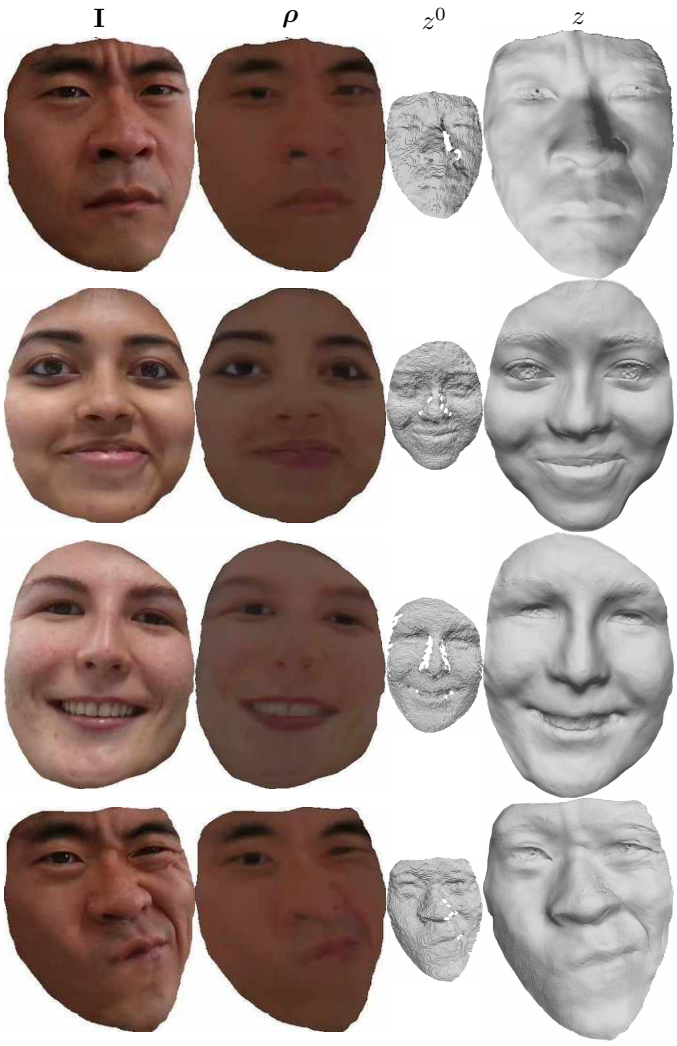


Fig. 5: Faces of these subjects were captured with an Intel Realsense D415 camera. Subjects were captured at $640 \times 480\ \mathrm{px}^2$ depth and $1280 \times 720\ \mathrm{px}^2$ RGB resolutions.

Figure 6 shows the case where the underlying geometry (using "Augustus" of [87]) represents a face, but different color and appearance in $\mathbf{I}$ lead to a wrong estimate of $\boldsymbol{\rho}$ of the deep net, resulting in artefacts on the depth estimate z during optimisation. Circumventing those issues can be done by acquiring more data in a photometric stereo manner, as we will discuss in the next section.

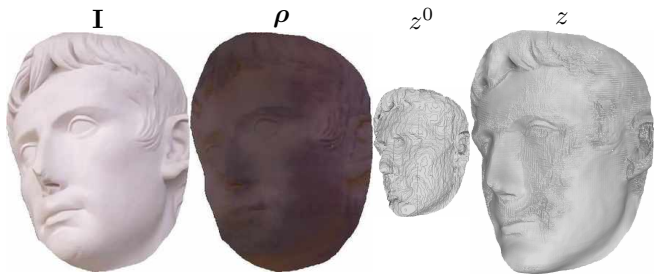


Fig. 6: Deep net failure case: A substantial departure from usual facial color and appearance results in incorrect albedo estimation of the deep net due to which the SfS optimisation fails and produces artefacts in the depth map.

5 Multi-Shot Depth Super-Resolution using Photometric Stereo

Performing photometric depth super-resolution from a single RGB-D frame requires some prior knowledge of the surface reflectance, either in terms of a piecewise-constantness prior or of adequation to the learning database. The only way to entirely get rid of such priors consists in acquiring multiple observations of the surface under varying lighting, i.e., performing uncalibrated photometric stereo to resolve depth super-resolution.

Let us thus consider from now on a sequence of images $\{\mathbf{I}_i\}_{i\in\mathcal{I}}$, with $\mathcal{I} := \{1,\dots,n\}$ and $n \geq 4$, captured under varying lighting conditions denoted by $\{\mathbf{l}_i\}_{i\in\mathcal{I}}$. The image formation model (4) is then turned into the following system of n equations:

$$\mathbf{I}^i = \boldsymbol{\rho}\,\langle \mathbf{m}_{z,\nabla z}, \mathbf{l}_i\rangle + \eta_I^i, \quad i \in \mathcal{I}. \tag{22}$$

In Eq. (22), neither the depth z nor the reflectance map $\boldsymbol{\rho}$ depend on i. Hence, their estimation is much more constrained in comparison with shape-from-shading. Nevertheless, nescience of the lighting vectors $\{\mathbf{l}_i\}_{i\in\mathcal{I}}$ makes the joint estimation of shape, reflectance and lighting i.e., uncalibrated photometric stereo, an ill-posed problem. As discussed in Section 2, the arising ambiguities cannot be resolved without the introduction of additional priors. As we shall see, in the context of RGB-D sensing the need for such priors can be circumvented and a purely data-driven approach can be followed. In other words, the low-resolution depth information act as a natural disambiguation prior for uncalibrated photometric stereo and, equally, the tailored photometric based-prior implicitly ensures surface regularity for depth map super-resolution.

5.1 Maximum Likelihood-Based Solution

Let us recall that the single-shot approach discussed in Section 3 required priors on the regularity of both the depth and the reflectance maps. By considering *multiple* RGB-D frames $\{\mathbf{I}_i, z_i^0\}_{i\in\mathcal{I}}$ of a static scene obtained under varying, yet unknown, lighting, we hope to end up with a variational framework free of such man-made priors. To this end, we consider a maximum likelihood framework instead of a Bayesian one.

Considering again the independence of depth and image observations as well as the independence of shape from reflectance and lighting, the joint likelihood of the observations $\{\mathbf{I}_i, z_i^0\}_{i\in\mathcal{I}}$ can be factored out as follows:

$$\begin{aligned}\mathcal{P}(\{\mathbf{I}_i, z_i^0\}_{i\in\mathcal{I}}|z,\boldsymbol{\rho},\{\mathbf{l}_i\}_{i\in\mathcal{I}}) = \\ \mathcal{P}(\{\mathbf{I}_i\}_{i\in\mathcal{I}}|z,\boldsymbol{\rho},\{\mathbf{l}_i\}_{i\in\mathcal{I}})\,\mathcal{P}(\{z_i^0\}_{i\in\mathcal{I}}|z).\end{aligned} \tag{23}$$

Under the assumption that the random variables η_I^i, $i \in \mathcal{I}$ in Eq. (22) are homoskedastically distributed according to zero-mean Gaussian laws with the same covariance matrix $\mathrm{diag}(\sigma_I^2,\sigma_I^2,\sigma_I^2)$, the marginal likelihood for $\{\mathbf{I}_i\}_{i\in\mathcal{I}}$ can be explicitly written as

$$\begin{aligned}\mathcal{P}(\{\mathbf{I}_i\}_{i\in\mathcal{I}}|z,\boldsymbol{\rho},\{\mathbf{l}_i\}_{i\in\mathcal{I}}) \propto \\ \exp\left\{-\frac{\sum_{i\in\mathcal{I}}\left\|\boldsymbol{\rho}\,\langle \mathbf{m}_{z,\nabla z},\mathbf{l}_i\rangle - \mathbf{I}_i\right\|^2_{\ell^2(\Omega_{HR})}}{2\sigma_I^2}\right\}.\end{aligned} \tag{24}$$

Assuming that the n low-resolution depth maps $\{z_i^0\}_{i\in\mathcal{I}}$ are consistent with the super-resolution model (1), and that the n corresponding random variables follow a zero-mean Gaussian distribution with same variance σ_z^2, the marginal likelihood for $\{z_i^0\}_{i\in\mathcal{I}}$ writes as

$$\mathcal{P}(\{z_i^0\}_{i\in\mathcal{I}}|z) \propto \exp\left\{-\frac{\sum_{i\in\mathcal{I}}\left\|Kz - z_i^0\right\|^2_{\ell^2(\Omega_{HR})}}{2\sigma_z^2}\right\}. \tag{25}$$

Maximum likelihood estimation of depth, reflectance and lighting consists in maximising the joint likelihood (23) or, equivalently, minimising its negative logarithm. Neglecting all additive constants and plugging (24) and (25) into (23), this writes as the following variational problem:

$$\begin{aligned}\min_{\substack{z:\Omega_{HR}\to\mathbb{R}\\ \boldsymbol{\rho}:\Omega_{HR}\to\mathbb{R}^3\\ \{\mathbf{l}_i\}_{i\in\mathcal{I}}\in\mathbb{R}^4}} \sum_{i\in\mathcal{I}} \Big\{ &\left\|Kz - z_i^0\right\|^2_{\ell^2(\Omega_{LR})} \\ &+\gamma\left\|\boldsymbol{\rho}\,\langle \mathbf{m}_{z,\nabla z},\mathbf{l}_i\rangle - \mathbf{I}_i\right\|^2_{\ell^2(\Omega_{HR})}\Big\},\end{aligned} \tag{26}$$

with the trade-off parameter γ given by the ratio $\gamma = \frac{\sigma_z^2}{\sigma_I^2}$. Let us emphasise the simplicity of the photometric stereo-based variational model (26), in comparison with the one obtained using shape-from-shading, cf. (11). Although one may think that more data introduces more complexity to such problems, we can clearly see here that in fact Problem (26) is naturally easier by itself as it does not include nonsmooth prior terms on the albedo and the depth, but only two data terms. As discussed next, this allows a much simpler numerical strategy to be followed.

5.2 Numerical Solving of (26)

Contrarily to the shape-from-shading problem (11), in (26) the nonlinearity arises only from the unit-length constraint on the normals. Therefore, we opt for a simpler numerical solution based on fixed point iterations. Considering Eqs (2) and (4), (26) can be rewritten as

$$\begin{aligned}\min_{\substack{z:\Omega_{HR}\to\mathbb{R}\\ \boldsymbol{\rho}:\Omega_{HR}\to\mathbb{R}^3\\ \{\mathbf{l}_i\}_{i\in\mathcal{I}}\in\mathbb{R}^4}} \sum_{i\in\mathcal{I}} \Big\{ &\left\|Kz - z_i^0\right\|^2_{\ell^2(\Omega_{LR})} \\ &+\gamma\left\|\frac{1}{d_{z,\nabla z}}\boldsymbol{\rho}\,\langle \tilde{\mathbf{m}}_{z,\nabla z},\mathbf{l}_i\rangle - \mathbf{I}_i\right\|^2_{\ell^2(\Omega_{HR})}\Big\}\end{aligned} \tag{27}$$

where $d_{z,\nabla z}$ is a scalar field ensuring the unit-length constraint for the normals:

$$d_{z,\nabla z} = \sqrt{|f\,\nabla z|^2 + \left(-z - \mathbf{p}\cdot\nabla z\right)^2}, \tag{28}$$

and $\tilde{\mathbf{m}}_{z,\nabla z}$ is a vector field encoding the normal direction:

$$\tilde{\mathbf{m}}_{z,\nabla z} = \begin{bmatrix} f\,\nabla z \\ -z - \mathbf{p}\cdot\nabla z \\ 1\end{bmatrix}. \tag{29}$$

In (27), only $d_{z,\nabla z}$ depends in a nonlinear way on the unknown depth z. Therefore, it seems natural to solve (27) iteratively, while freezing the nonlinearity[4]. At iteration (k)

4. Contrarily to the shape-from-shading case, in photometric stereo we found this fixed point strategy to be convergent. We however leave the convergence proof for future work.

and with the current estimates $(\boldsymbol{\rho}^{(k)}, \{\mathbf{l}_i^{(k)}\}_{i\in\mathcal{I}}, z^{(k)})$, one sweep of this scheme reads:

$$\boldsymbol{\rho}^{(k+1)} = \underset{\boldsymbol{\rho}:\Omega_{HR}\to\mathbb{R}^3}{\arg\min} \sum_{i\in\mathcal{I}} \left\| \frac{\left\langle \tilde{\mathbf{m}}_{z^{(k)},\nabla z^{(k)}}, \mathbf{l}_i^{(k)} \right\rangle}{d_{z^{(k)},\nabla z^{(k)}}} \boldsymbol{\rho} - \mathbf{I}_i \right\|^2_{\ell^2(\Omega_{HR})}, \quad (30)$$

$$\mathbf{l}_i^{(k+1)} = \underset{\mathbf{l}_i\in\mathbb{R}^4}{\arg\min} \left\| \frac{\boldsymbol{\rho}^{(k+1)}}{d_{z^{(k)},\nabla z^{(k)}}} \left\langle \tilde{\mathbf{m}}_{z^{(k)},\nabla z^{(k)}}, \mathbf{l}_i \right\rangle - \mathbf{I}_i \right\|^2_{\ell^2(\Omega_{HR})} \quad \forall i, \quad (31)$$

$$\begin{aligned} z^{(k+1)} = \underset{z:\Omega_{HR}\to\mathbb{R}}{\arg\min} & \sum_{i\in\mathcal{I}} \left\| Kz - z_i^0 \right\|^2_{\ell^2(\Omega_{LR})} \quad (32) \\ & + \gamma \left\| \frac{\boldsymbol{\rho}^{(k+1)}}{d_{z^{(k)},\nabla z^{(k)}}} \left\langle \tilde{\mathbf{m}}_{z,\nabla z}, \mathbf{l}_i^{(k+1)} \right\rangle - \mathbf{I}_i \right\|^2_{\ell^2(\Omega_{HR})}. \end{aligned}$$

All three problems (30), (31) and (32) are linear least-squares problems which we solve using the conjugate gradient method on the normal equations.

Our initial values for $(k) = (0)$ are chosen to be $\boldsymbol{\rho}^{(0)} = \text{mean}(\{\mathbf{I}_i\}_{i\in\mathcal{I}})$, $\mathbf{l}_i^{(0)} = [0, 0, -1, 0]^\top \ \forall i$, $z^{(0)}$ a smoothed version of $\text{mean}(\{z_i^0\}_{i\in\mathcal{I}})$ using the guided filter [74] followed by the bicubic interpolation to upsample to the image domain Ω_{HR}, $\boldsymbol{\theta}^{(0)} = (z, \nabla z)^{(0)}$, $\mathbf{u}^{(0)} = 0$. Here mean denotes the averaging operator. As in Section 3.2, to verify convergence we check if the relative residual r_{rel} falls below some threshold. The overall scheme is implemented in Matlab. The whole process lasts, depending on the dataset and number of images between $30s$ and $360s$, see the experiments in Figure 16 in supplementary material.

5.3 Experiments

This section is dedicated to the experiments on the joint approach of depth super-resolution and photometric stereo described above. We use synthetic and real-world data to support the claim of being able to do detail-preserving depth super-resolution under no albedo constraint.

5.3.1 Synthetic Data

Synthetic data is generated in a similar manner as described in Section 3.3. To this end we use the datasets of [88], [89] and render the meshes to depth maps of resolutions of $[60\times80\ \text{px}^2]$, $[120\times160\ \text{px}^2]$, $[240\times320\ \text{px}^2]$, $[640\times480\ \text{px}^2]$. The highest resolution depth map is then used to generate RGB images with complex Lambertian reflectance under the spherical harmonics model. The light vectors $\{\mathbf{l}_i\}_{i\in\mathcal{I}}$ are generated randomly. Finally zero-mean Gaussian noise of is added to the high resolution RGB and low-resolution depth image. For fair comparison $\sigma_{\{I,z\}}$ are chosen in the same manner as in Section 3.3. Figure 15 in the supplementary material summarises the resulting RGB images.

Number of Images We first evaluate the impact of the number of input images on the RMSE and MAE, as well as the runtime. Figure 16 in supplementary material shows clearly that more images result in better depth reconstruction, yet with linear increasing runtime. Thus, we can say that $n \in [10, 30]$ seems to be resulting in good reconstructions without increasing the runtime extraordinarily. All experiments were run on a machine with a CPU with $3.5GHz$ and $16GB$ of RAM. Convergence was reached within at most 15 iterations in all our experiments.

Parameter Tuning The parameter γ in (26) is the only tuning parameter for the joint approach of depth map super-resolution and uncalibrated photometric stereo. Although it can be deduced from the statitics, we consider it a hyper parameter, similarly to Section 3.3. From (26) we can nicely deduce that for $\gamma \to 0$ yields a pure regularisation-free depth super-resolution scheme which can be solved pointwise, thus the resulting depth map is as noisy as the input, cf. Figure 17 in the supplementary material. On the other hand if $\gamma \to \infty$ the resulting variational problem is a regularisation-free uncalibrated photometric stereo scheme, i.e. the loss of low-frequency information from the depth prior introduces a high RMSE value, due to the Lorentz ambiguity. Combining both ill-posed problems, namely depth super-resolution and uncalibrated photometric stereo, by incorporating them in a variational scheme with a trade-off parameter $\xi \in [10^{-2}, 10^{1}]$ appears to result in low RMSE and MAE values.

Quantitative Comparison Table 3 and Figure 18 in the supplementary material show the quantitative results of other methods and our approach on our synthetic dataset. Our approach outperforms state-of-the-art methods such as pure uncalibrated photometric stereo [41], shading-based depth refinement using a low-resolution RGB image [51] and image-driven depth super-resoution using an anisotropic Huber-loss as regularisation term [1], [90]. Latter assumes large gradients in the RGB image imply larger gradients in the depth image. The method of [41] estimates albedo along with the geometry and lighting, but it uses no low-resolution depth clues in order to resolve the general bas-relief ambiguity. RGB-D fusion [51] makes refines a noisy depth map using shape-from-shading, but at low-resolution and using a single image. As we saw in Section 3 and 4, shape-from-shading methods need a regularisation term on the albedo, which introduces man-made assumptions and if these are not met the resulting depth shows artefacts of propagated reflectance information. Only a photometric stereo setup can avoid resorting to a reflectance prior, and in combination with low-resolution depth clues it provides the best results with no photometric ambiguity.

5.3.2 Real-World Data

RGB-D photometric stereo real-world datasets are captured using either the Asus Xtion Pro Live or Intel Realsense D415 camera. To acquire data under various resolutions, we consider RGB images of resolution $[1280 \times 1024\ \text{px}^2]$ and depth images of resolution QVGA ($[320 \times 240\ \text{px}^2]$) or VGA ($[640 \times 480\ \text{px}^2]$) for the former camera, while $[1280\times720\ \text{px}^2]$ is the RGB resolution and VGA is depth resolution for the latter. The setup is the same as in Section 3.3, just multiple images of the same static scene with static camera under varying lighting conditions are captured. Varying lighting conditions can be realised using a handheld LED light source, moving it freely around to illuminate the scene from different directions during the capturing process. As we consider an uncalibrated setup, these images can be easily captured. From each image sequence we extract $n = 20$ random high-resolution RGB images $\mathbf{I}_i$ and low-resolution depth images z_i^0.

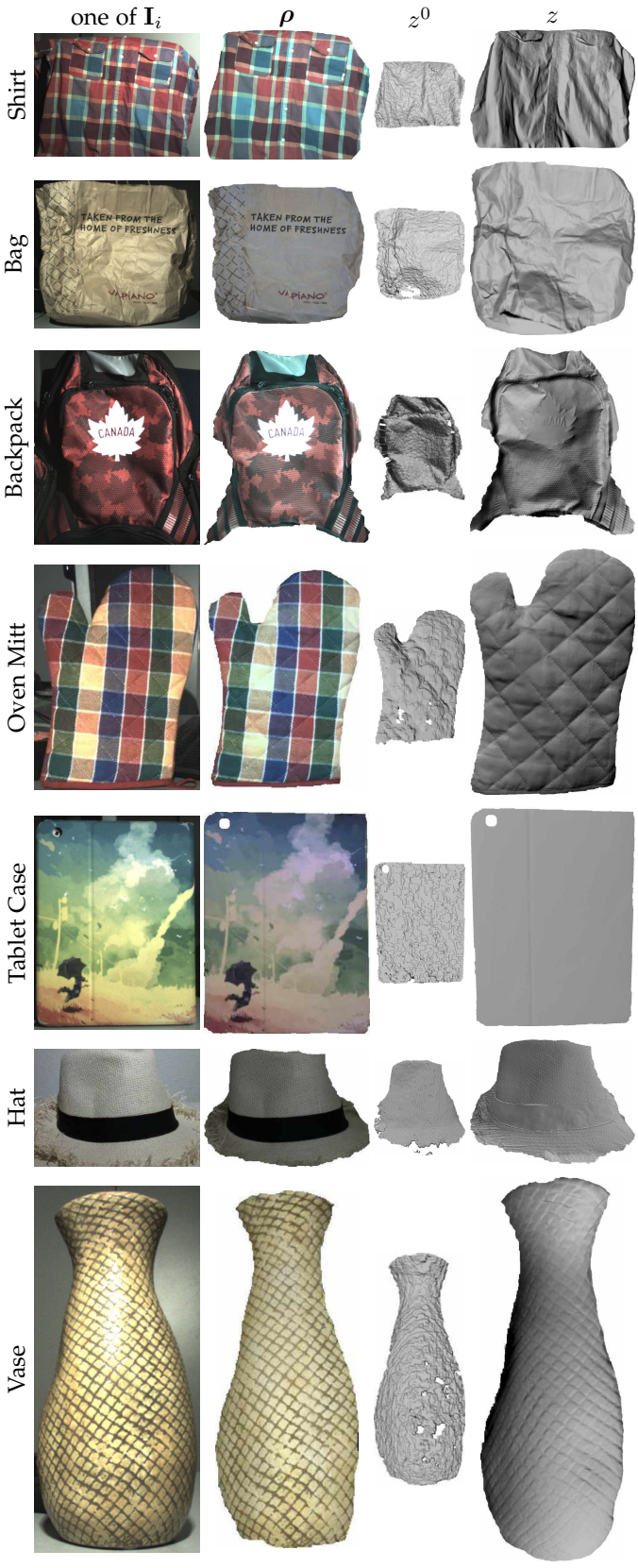


Fig. 7: Qualitative results on real-world datasets using our uncalibrated photometric stereo method. "Hat" was captured using the RealSense D415. The remaining datasets were captured with an Asus Xtion Pro Live RGB-D camera.

The results can be seen in Figure 7. We consider seven different objects: "Shirt", "Bag", "Backpack", "Oven Mitt", "Tablet Case", "Hat" and "Vase". Each dataset consists of challenging reflectance maps as well as fine-scale geometric details. For the "Shirt" dataset the fine wrinkles are nicely recovered and even the geometry of the buttons is visible in the estimated depth z. The "Bag" dataset comprises of difficult geometry and a purely reflectance-based text printing, our method is able to separate the geometric information from the reflectance information and thus none of the printed text is propagated to geometry. The "Backpack" has very thin geometric structures which are nicely recovered and although the reflectance is partially very low, it does not seem to deteriorate the depth estimate. This is probably due to having a rough prior on shape. The "Oven Mitt" consists of fine stichting structures which are nicely recovered and separated from the estimated reflectance. The "Tablet Case" consists of a very complex smoothly varying albedo with a fine groove and it can be seen that almost no reflectance properties are propagated to the geometry, but the groove is still visible. Nevertheless, it is also still visible in the reflectance estimate, yet this does not seem to affect the depth estimate strongly. The "Hat" consists of a simpler albedo, but with very fine geometric details which are nicely recovered in the depth estimate, although it seems that shading information is visible in the reflectance estimate. Albeit part of the reflectance is low, we may note that this does not seem to affect the resulting depth in a negative way, thanks to the low-resolution depth prior. Interestingly, although our method is based on the Lambertian reflectance assumption embedded in the spherical harmonics approximation, the high-quality shape of the reflective "Vase" can still be reconstructed and even where color is saturated at the specular regions, fine scale geometric details are nicely visible. This may be due to the reason that we use multiple images under varying light conditions and the specularities in a certain image may not exist in others. Thus, the reflections will not affect the shape estimation much.

Clearly, the advantage of having no albedo regularisation can be noticed as the estimation of even very difficult reflectance combined with fine geometric structures is no problem. More qualitative comparisons against other state-of-the-art methods can be found in the supplementary material.

6 Conclusion

We put forward several photometric approaches for solving the depth super-resolution problem in RGB-D sensing. All of them can be used out-of-the-box with any commodity RGB-D camera without the need of tedious calibration.

First a single-shot approach based on shape-from-shading was proposed. The low-resolution depth cues resolve the ambiguities arising in shape-from-shading and symmetrically, high-resolution photometric cues resolve those of depth super-resolution. Albeit the results were promising, they remained limited by piecewise-constant albedo assumption.

To overcome this issue we suggested to loosen this assumption by introducing a deep neural network to estimate the albedo. This allowed us to handle arbitrarily complex

albedo, as long as the object of interest falls into the training set of the deep network, e.g. faces.

To completely bypass the need for any kind of prior on the albedo, we moved from a single-shot setup to a multi-shot setup. The latter allowed us to simultaneously solve uncalibrated photometric stereo and depth super-resolution. This purely data-driven approach requires no man-made regularisation term or learning database, at the cost of acquiring more data.

As future work, we will explore the theoretical properties of the proposed approaches and prove uniqueness of the solutions by resorting to a continuous analysis of the problem.

ACKNOWLEDGMENTS

The authors wish to thank Thomas Möllenhoff for helpful discussions and comments, also they would like to thank Robert Maier for code to successfully render `ply`-files to depth maps.

REFERENCES

[1] M. Unger, T. Pock, M. Werlberger, and H. Bischof, "A convex approach for variational super-resolution," in *Joint Pattern Recognition Symposium*, 2010, pp. 313–322.

[2] J. Park, H. Kim, Y.-W. Tai, M. S. Brown, and I. S. Kweon, "High quality depth map upsampling for 3F-TOF cameras," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 1623–1630.

[3] Y. Quéau, J.-D. Durou, and J.-F. Aujol, "Normal Integration: A Survey," *Journal of Mathematical Imaging and Vision*, vol. 60, no. 4, pp. 576–593, 2018.

[4] 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 the European Conference on Computer Vision (ECCV)*, 2004, pp. 574–587.

[5] R. Basri and D. P. Jacobs, "Lambertian reflectances and linear subspaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 2, pp. 218–233, 2003.

[6] R. Ramamoorthi and P. Hanrahan, "An Efficient Representation for Irradiance Environment Maps," in *Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques*, 2001, pp. 497–500.

[7] B. Haefner, Y. Quéau, T. Möllenhoff, and D. Cremers, "Fight ill-posedness with ill-posedness: Single-shot variational depth super-resolution from shading," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 164–174.

[8] S. Peng, B. Haefner, Y. Quéau, and D. Cremers, "Depth super-resolution meets uncalibrated photometric stereo," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, 2017, pp. 2961–2968.

[9] B. Goldlücke, M. Aubry, K. Kolev, and D. Cremers, "A super-resolution framework for high-accuracy multiview reconstruction," *International Journal of Computer Vision*, vol. 106, no. 2, pp. 172–191, 2014.

[10] R. Maier, J. Stückler, and D. Cremers, "Super-resolution keyframe fusion for 3D modeling with high-quality textures," in *Proceedings of the International Conference on 3D Vision (3DV)*, 2015, pp. 536–544.

[11] S. Schuon, C. Theobalt, J. Davis, and S. Thrun, "Lidarboost: Depth superresolution for TOF 3D shape scanning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 343–350.

[12] O. Mac Aodha, N. D. F. Campbell, A. Nair, and G. J. Brostow, "Patch based synthesis for single depth image super-resolution," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2012, pp. 71–84.

[13] J. Xie, R. S. Feris, and M.-T. Sun, "Edge-guided single depth image super resolution," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 428–438, 2016.

[14] M. Hornácek, C. Rhemann, M. Gelautz, and C. Rother, "Depth super resolution by rigid body self-similarity in 3D," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 1123–1130.

[15] J. Li, Z. Lu, G. Zeng, R. Gan, and H. Zha, "Similarity-aware patchwork assembly for depth image super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 3374–3381.

[16] J. Xie, R. S. Feris, S.-S. Yu, and M.-T. Sun, "Joint super resolution and denoising from a single depth image," *IEEE Transactions on Multimedia*, vol. 17, no. 9, pp. 1525–1537, 2015.

[17] D. Ferstl, M. Rüther, and H. Bischof, "Variational depth superresolution using example-based edge representations," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 513–521.

[18] G. Riegler, M. Rüther, and H. Bischof, "ATGV-net: accurate depth super-resolution," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016, pp. 268–284.

[19] B. K. P. Horn, "Shape From Shading: A Method for Obtaining the Shape of a Smooth Opaque Object From One View," Ph.D. dissertation, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 1970.

[20] 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*, vol. 5, no. 1, pp. 311–342, 2012.

[21] J.-D. Durou, M. Falcone, and M. Sagona, "Numerical Methods for Shape-from-shading: A New Survey with Benchmarks," *Computer Vision and Image Understanding*, vol. 109, no. 1, pp. 22–43, 2008.

[22] R. Zhang, P.-S. Tsai, J. E. Cryer, and M. Shah, "Shape-from-shading: a survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 8, pp. 690–706, 1999.

[23] B. K. P. Horn and M. J. Brooks, "The variational approach to shape from shading," *Computer Vision, Graphics, and Image Processing*, vol. 33, no. 2, pp. 174–208, 1986.

[24] K. Ikeuchi and B. K. Horn, "Numerical shape from shading and occluding boundaries," *Artificial intelligence*, vol. 17, no. 1-3, pp. 141–184, 1981.

[25] E. Cristiani and M. Falcone, "Fast semi-lagrangian schemes for the eikonal equation and applications," *SIAM Journal on Numerical Analysis*, vol. 45, no. 5, pp. 1979–2011, 2007.

[26] M. Falcone and M. Sagona, "An algorithm for the global solution of the shape-from-shading model," in *Proceedings of the International Conference on Image Analysis and Processing (ICIAP)*, 1997, pp. 596–603.

[27] P.-L. Lions, E. Rouy, and A. Tourin, "Shape-from-shading, viscosity solutions and edges," *Numerische Mathematik*, vol. 64, no. 1, pp. 323–353, 1993.

[28] E. Rouy and A. Tourin, "A viscosity solutions approach to shape-from-shading," *SIAM Journal on Numerical Analysis*, vol. 29, no. 3, pp. 867–884, 1992.

[29] A. R. Bruss, "The eikonal equation: Some results applicable to computer vision," *Journal of Mathematical Physics*, vol. 23, no. 5, pp. 890–896, 1982.

[30] E. H. Adelson and A. P. Pentland, *Perception as Bayesian inference*. Cambridge University Press, 1996, ch. The perception of shading and reflectance, pp. 409–423.

[31] R. Huang and W. A. P. Smith, "Shape-from-shading under complex natural illumination," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2011, pp. 13–16.

[32] M. K. Johnson and E. H. Adelson, "Shape estimation in natural illumination," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 2553–2560.

[33] S. R. Richter and S. Roth, "Discriminative shape from shading in uncalibrated illumination," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1128–1136.

[34] Y. Quéau, J. Mélou, F. Castan, D. Cremers, and J.-D. Durou, "A Variational Approach to Shape-from-shading Under Natural Illumination," in *Energy Minimization Methods for Computer Vision and Pattern Recognition (EMMCVPR)*, 2017, pp. 342–357.

[35] J. Barron and J. Malik, "Shape, illumination, and reflectance from shading," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 8, pp. 1670–1687, 2015.

[36] R. J. Woodham, "Photometric Method for Determining Surface Orientation from Multiple Images," *Optical Engineering*, vol. 19, no. 1, pp. 139–144, 1980.

[37] H. Hayakawa, "Photometric stereo under a light source with arbitrary motion," *Journal of the Optical Society of America A*, vol. 11, no. 11, pp. 3079–3089, 1994.

[38] P. N. Belhumeur, D. J. Kriegman, and A. L. Yuille, "The bas-relief ambiguity," *International Journal of Computer Vision*, vol. 35, no. 1, pp. 33–44, 1999.

[39] R. Basri, D. W. Jacobs, and I. Kemelmacher, "Photometric stereo with general, unknown lighting," *International Journal of Computer Vision*, vol. 72, no. 3, pp. 239–257, 2007.

[40] N. G. Alldrin, S. P. Mallick, and D. J. Kriegman, "Resolving the generalized bas-relief ambiguity by entropy minimization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[41] T. Papadhimitri and P. Favaro, "A closed-form, consistent and robust solution to uncalibrated photometric stereo via local diffuse reflectance maxima," *International Journal of Computer Vision*, vol. 107, no. 2, pp. 139–154, 2014.

[42] Y. Quéau, F. Lauze, and J.-D. Durou, "Solving Uncalibrated Photometric Stereo using Total Variation," *Journal of Mathematical Imaging and Vision*, vol. 52, no. 1, pp. 87–107, 2015.

[43] F. Lu, X. Chen, I. Sato, and Y. Sato, "Symps: Brdf symmetry guided photometric stereo for shape and light source estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 1, pp. 221–234, 2018.

[44] Y. Quéau, T. Wu, F. Lauze, J.-D. Durou, and D. Cremers, "A Non-Convex Variational Approach to Photometric Stereo under Inaccurate Lighting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 350–359.

[45] 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*, vol. 122, no. 1, pp. 1–16, 2017.

[46] R. Maier, K. Kim, D. Cremers, J. Kautz, and M. Nießner, "Intrinsic3d: High-quality 3D reconstruction by joint appearance and geometry optimization with spatially-varying lighting," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3114–3122.

[47] 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*, vol. 34, no. 4, pp. 96:1–96:14, 2015.

[48] 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 the IEEE International Conference on Computer Vision (ICCV)*, 2013, pp. 1617–1624.

[49] K. Kim, A. Torii, and M. Okutomi, "Joint estimation of depth, reflectance and illumination for depth refinement," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 199–207.

[50] 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 the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4378–4386.

[51] R. Or-El, G. Rosman, A. Wetzler, R. Kimmel, and A. Bruckstein, "RGBD-Fusion: Real-Time High Precision Depth Recovery," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 5407–5416.

[52] 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*, vol. 33, no. 6, pp. 200:1–200:10, 2014.

[53] L.-F. Yu, S.-K. Yeung, Y.-W. Tai, and S. Lin, "Shading-based shape refinement of RGB-D images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 1415–1422.

[54] R. Anderson, B. Stenger, and R. Cipolla, "Augmenting depth camera output using photometric stereo," in *Proceedings of the IAPR Conference on Machine Vision Applications (MVA)*, 2011, pp. 369–372.

[55] A. Chatterjee and V. Madhav Govindu, "Photometric refinement of depth maps for multi-albedo objects," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 933–941.

[56] L. Xie, Y. Xu, X. Zhang, W. Bao, C. Tong, and B. Shi, "A self-calibrated photo-geometric depth camera," *The Visual Computer*, 2018.

[57] Y. Zhang, Q. Zhang, and W. Feng, "High-Resolution Depth Refinement by Photometric and Multi-shading Constraints," in *PRICAI 2018: Trends in Artificial Intelligence*, 2018, pp. 201–209.

[58] J. Diebel and S. Thrun, "An application of Markov random fields to range sensing," in *Advances in Neural Information Processing Systems*, 2006, pp. 291–298.

[59] D. Ferstl, C. Reinbacher, R. Ranftl, M. Rüther, and H. Bischof, "Image guided depth upsampling using anisotropic total generalized variation," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013, pp. 993–1000.

[60] Q. Yang, R. Yang, J. Davis, and D. Nistér, "Spatial-depth super resolution for range images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[61] B. Li, Y. Zhou, Y. Zhang, and A. Wang, "Depth image super-resolution based on joint sparse coding," *Pattern Recognition Letters*, 2018.

[62] P. Tan, S. Lin, and L. Quan, "Subpixel photometric stereo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 8, pp. 1460–1471, 2008.

[63] S. Chaudhuri and M. V. Joshi, *Motion-free super-resolution*. Springer Verlag, 2005.

[64] Z. Lu, Y.-W. Tai, F. Deng, M. Ben-Ezra, and M. S. Brown, "A 3D imaging framework based on high-resolution photometric-stereo and low-resolution depth," *International Journal of Computer Vision*, vol. 102, no. 1-3, pp. 18–32, 2013.

[65] D. Mumford, "Bayesian rationale for the variational formulation," in *Geometry-driven diffusion in computer vision*, 1994, pp. 135–146.

[66] G. Graber, J. Balzer, S. Soatto, and T. Pock, "Efficient minimal-surface regularization of perspective depth maps in variational stereo," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 511–520.

[67] E. H. Land, "The retinex theory of color vision," *Scientific American*, vol. 237, no. 6, pp. 108–120, 1977.

[68] 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*, vol. 3, no. 1, pp. 1–122, 2011.

[69] J. Eckstein and D. P. Bertsekas, "On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Mathematical Programming*, vol. 55, no. 1, pp. 293–318, 1992.

[70] R. Glowinski and A. Marroco, "Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de Dirichlet non linéaires," *Revue française d'automatique, informatique, recherche opérationnelle. Analyse numérique*, vol. 9, no. R2, pp. 41–76, 1975.

[71] E. Strekalovskiy and D. Cremers, "Real-time minimization of the piecewise smooth Mumford-Shah functional," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014, pp. 127–141.

[72] M. Schmidt, "minFunc: unconstrained differentiable multivariate optimization in Matlab," 2005, http://www.cs.ubc.ca/~schmidtm/Software/minFunc.html.

[73] D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Mathematical programming*, vol. 45, no. 1, pp. 503–528, 1989.

[74] K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 6, pp. 1397–1409, 2013.

[75] K. Khoshelham and S. O. Elberink, "Accuracy and resolution of Kinect depth data for indoor mapping applications," *Sensors*, vol. 12, no. 2, pp. 1437–1454, 2012.

[76] W.-C. Ma, T. Hawkins, P. Peers, C.-F. Chabert, M. Weiss, and P. Debevec, "Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination," in *Proceedings of the 18th Eurographics Conference on Rendering Techniques*, 2007, pp. 183–194.

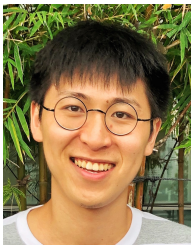
[77] G. Stratou, A. Ghosh, P. Debevec, and L. Morency, "Effect of illumination on automatic expression recognition: A novel 3d relightable facial database," in *Face and Gesture*, 2011, pp. 611–618.

[78] J. Shen, X. Yang, Y. Jia, and X. Li, "Intrinsic images using optimization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 3481–3487.

[79] Q. Fan, J. Yang, G. Hua, B. Chen, and D. Wipf, "Revisiting deep intrinsic image decompositions," in *Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8944–8952.

[80] C. Li, K. Zhou, and S. Lin, "Intrinsic face image decomposition with human face priors," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014, pp. 218–233.
[81] G. Trigeorgis, P. Snape, I. Kokkinos, and S. Zafeiriou, "Face normals "in-the-wild" using fully convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 38–47.
[82] Z. Shu, E. Yumer, S. Hadap, K. Sunkavalli, E. Shechtman, and D. Samaras, "Neural face editing with intrinsic image disentangling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5444–5453.
[83] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y.-G. Jiang, "Pixel2mesh: Generating 3d mesh models from single rgb images," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.
[84] S. Sengupta, A. Kanazawa, C. D. Castillo, and D. W. Jacobs, "SfSNet: Learning Shape, Refectance and Illuminance of Faces in the Wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6296–6305.
[85] J. Shi, Y. Dong, H. Su, and S. X. Yu, "Learning non-lambertian object intrinsics across shapenet categories," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5844–5853.
[86] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, pp. 234–241.
[87] M. Zollhöfer, A. Dai, M. Innman, C. Wu, M. Stamminger, C. Theobalt, and M. Nießner, "Shading-based Refinement on Volumetric Signed Distance Functions," 2015, http://graphics.stanford.edu/projects/vsfs/.
[88] M. Levoy, J. Gerth, B. Curless, and K. Pull, "The stanford 3d scanning repository," 2005, http://www-graphics.stanford.edu/data/3dscanrep.
[89] "The joyful yell," 2015, https://www.thingiverse.com/thing:897412.
[90] M. Werlberger, W. Trobin, T. Pock, A. Wedel, D. Cremers, and H. Bischof, "Anisotropic Huber-L1 Optical Flow," in *Proceedings of the British Machine Vision Conference*, 2009, pp. 108.1–108.11.

Bjoern Haefner received his B.Sc. in Mathematics from the OTH Regensburg in 2013 and his M.Sc. in Mathematics in Science and Engineering from the Technical University of Munich in 2016. Since mid November 2016, he is a full-time PhD student in the Computer Vision and Artificial Intelligence chair at the Technical University of Munich. His research interests include RGB-D data processing for 3D reconstruction using variational methods.



Songyou Peng received the Erasmus Mundus M.Sc. in Computer Vision and Robotics in 2017. Between 2016 and 2017, he spent some time doing research at INRIA Grenoble and Technical University of Munich. Since 2018 he is a research engineer at Advanced Digital Sciences Center in Singapore. His research interests are computer vision and machine learning.

Alok Verma is pursuing a Master's degree in Biomedical Computing at the Technical University of Munich, Germany since 2017. Previously he worked as a senior electrical and software engineer at Philips Healthcare, Bangalore, India focusing on C-Arm X-ray Systems. His research interests are computer vision and deep learning for medical and non-medical images.

Yvain Quéau received his Ph.D from INP-ENSEEIHT, Université de Toulouse, in 2015. From 2016 to 2018 he was a postdoctoral researcher in Technical University Munich, Germany, and then an associate processor with ISEN Brest, France. Since 2018 he is a CNRS researcher with the GREYC laboratory, Université de Caen, France. His research focuses on variational methods for solving inverse problems in computer vision.

Daniel Cremers received the PhD degree in computer science from the University of Mannheim, Germany. Subsequently, he spent two years as a postdoctoral researcher with UCLA and one year as a permanent researcher at Siemens Corporate Research, Princeton. From 2005 until 2009, he was associate professor with the University of Bonn. Since 2009 he holds the Chair of Computer Vision and Artificial Intelligence at the Technical University of Munich. He received numerous awards including the Gottfried-Wilhelm Leibniz Award 2016, the biggest award in German academia.