Uncalibrated, Unified and Unsupervised Specular-aware Photometric Stereo

Pablo Estevez and Antonio Agudo
Institut de Robòtica i Informàtica Industrial, CSIC-UPC
Barcelona, 08028, Spain

***Abstract*—In this paper we present a variational approach to simultaneously recover the 3D reconstruction, reflectance, lighting and specularities of an object, all of them, from a set of RGB images. The approach works in an uncalibrated, unified and unsupervised manner, without assuming any prior knowledge of the shape geometry or training data to constrain the solution and under general lighting. To this end, the approach exploits a physically-aware image formation model that in combination with a perspective projection one and under spherical harmonics lighting gives a fully interpretable algorithm. Integrability is implicitly ensured as the shape is coded by a depth map rather than normal vectors. As a consequence, a wide variety of illumination conditions and complex geometries can be acquired. Our claims have been experimentally validated on challenging synthetic and real datasets, obtaining a good trade-off between accuracy and computational budget in comparison with competing approaches.**

I. Introduction

Photometric stereo (PS) is a relevant problem in computer vision and pattern recognition aiming to retrieve both the 3D geometry of an object –probably via its normal vectors– and the reflectance of a scene from multiple images taken at the same viewpoint but under different illumination conditions [23], [24], [25]. Frequently, the way to sort out the problem consists in inverting a physically-aware image formation model that assumes a certain control of lighting. Despite providing accurate formulations, lighting control reduces the applicability of these methods to laboratory scenarios where an exhaustive calibration of lighting must be performed. Instead, the uncalibrated counterpart is ill-posed as the underlying normal map is recovered up to a linear ambiguity [5], that decreases to a generalized bas-relief one if integrability is enforced. Differential approaches based on variational formulations were introduced in order to directly retrieve the 3D geometry as a depth map, relaxing the remaining ambiguities [2], [16]. The previous works assumed an illumination induced by a single source, limiting strongly their applicability to natural scenarios where general lighting conditions appear. A non-directional illumination model was proposed by [7] to encode natural illumination that was even used in cloudy days [8]. Unfortunately, solving PS for uncalibrated and general lighting is a very challenging task that has been rarely explored. Some exceptions were proposed [15], [19] by assuming coarse 3D reconstructions as a prior, spatially-varying directional lighting models [12], or Lambertian materials [4]. Beyond Lambertian PS, the problem can be even more challenging for non-Lambertian objects as its appearance could contain a mixture of specular and diffuse reflection properties.

Meth. \ Feat.	Integrability	Uncalibrated		Prior		Material	
		No	Yes	Shape	Training	L	NL
[8], [7]		✓				✓	
[15], [19]	✓		✓	✓		✓	
[12]			✓			✓	
[11], [22], [26]		✓			✓	✓	✓
[4]	✓		✓			✓	
Ours	✓		✓			✓	✓

TABLE I
Qualitative comparison of our approach with respect to competing methods. Our approach is the only one that directly retrieves a depth map (potentially non-integrable surface normals are not estimated) of both Lambertian (L) and non-Lambertian (NL) materials; it is uncalibrated and works under general lighting; and it assumes neither shape priors nor training data with ground truth.

In parallel, the use of deep learning has been also applied to PS [9], [11], [22], [26]. These approaches propose end-to-end supervised training methodologies that ignore the underlying physical principle of PS as the model is learned from data. While the obtained results are promising, their lack of physical interpretability prevents them from exploiting the real interactions between surface geometry and specularities, as the last are handled in an implicit manner. However, the previous has not prevented a wider up-taking of these methods. The lack of large-scale real-world training data hinders the progress of data-driven deep PS. It is worth noting that obtaining ground truth –including 3D geometries and lighting– in a natural lab setting is both expensive and laborious, especially for certain object materials where the acquisition could be very hard. In addition to that, these approaches normally assume the light direction as well as the intensity at each illumination instant, i.e., the problem is solved in a calibrated manner. Table I summarizes a qualitative comparison in terms of available characteristics and assumptions of our approach and the most relevant competing approaches. As it can be seen, our approach is the only one that has all characteristics without assuming strong priors such as the use of large amounts of training data or a 3D rough geometry.

In this paper we overcome most of the limitations of current methods with a variational algorithm that can solve the PS problem for non-specific objects. Our approach is unified, unsupervised –no ground truth is needed for supervision–, and efficient; as well as being able to handle the problem in an uncalibrated manner under general lighting.

II. Physically-aware Photometric Stereo

Let us consider a set of observations $\{\mathcal{I}_c^i \subset \mathbb{R}^2\}$ composed of $i = \{1, \dots, I\}$ images with $c = \{1, \dots, C\}$ color channels where it appears an object we want to reconstruct in 3D. For that object, we also define $\mathcal{M} \subset \mathcal{I}_c^i$ as its shape segmentation in the image set, i.e., the masked pixel domain. Considering that the object to be captured is Lambertian, the surface reflectance for all P pixel points $\mathbf{p} \in \mathcal{M}$ can be modeled by collecting elementary luminance contributions arising from all the incident lighting directions $\boldsymbol{\omega}$ as:

$$\mathcal{I}_c^i(\mathbf{p}) = \int_{\mathbb{H}^2} \rho_c(\mathbf{p})\, l_c^i(\boldsymbol{\omega}) \max\{0, \boldsymbol{\omega}^\top \mathbf{n}(\mathbf{p})\}\, \mathrm{d}\omega, \tag{1}$$

where $\mathbb{H}^2$ is the unit sphere in $\mathbb{R}^3$, $\rho_c(\mathbf{p})$ and $l_c^i(\boldsymbol{\omega})$ indicate the color-wise albedos and intensity of the incident lights, respectively, and $\mathbf{n}(\mathbf{p})$ the unit-length surface normals at the surface point conjugate to pixel $\mathbf{p}$. The Lambertian surface assumes $\rho_c(\mathbf{p})$ to be always a positive value. The object irradiance or shading component is coded by the max operator. Unfortunately, the previous model cannot handle materials that exhibit a combination of specular and diffuse reflection properties, i.e., non-Lambertian surfaces. To solve that, we can follow the approximation of a Phong reflection model that considers the light at the p-th pixel as the sum of two additive terms: a viewpoint-independent diffuse and a view-dependent specular. In other words, as the model in Eq. (1) is specularity-free diffuse, we consider an additive component $s^i(\mathbf{p})$ for the specular reflection as:

$$\mathcal{I}_c^i(\mathbf{p}) = \int_{\mathbb{H}^2} \rho_c(\mathbf{p})\, l_c^i(\boldsymbol{\omega}) \max\{0, \boldsymbol{\omega}^\top \mathbf{n}(\mathbf{p})\}\, \mathrm{d}\omega + s^i(\mathbf{p}). \tag{2}$$

According to literature [2], [23], [25], the PS problem in an uncalibrated manner consists in recovering the 3D shape of the object (via its normals $\mathbf{n}(\mathbf{p})$) together with the quantities $\{\rho_c\}$, $\{l_c^i\}$ and $\{s^i\}$, all of them, from the set $\{\mathcal{I}_c^i\}$.

Following [1] the irradiance map can be modeled using a spherical harmonic approximation of general lighting by means of a half-cosine kernel:

$$k(\boldsymbol{\omega}, \mathbf{n}(\mathbf{p})) = \max\{0, \boldsymbol{\omega}^\top \mathbf{n}(\mathbf{p})\}. \tag{3}$$

The general image formation model in Eq. (2) can now be written as:

$$\begin{aligned}\mathcal{I}_c^i(\mathbf{p}) =& \rho_c(\mathbf{p}) \int_{\mathbb{H}^2} l_c^i(\boldsymbol{\omega})\; k(\boldsymbol{\omega}, \mathbf{n}(\mathbf{p}))\; \mathrm{d}\omega + s^i(\mathbf{p}) \\ =& \rho_c(\mathbf{p}) \alpha(\boldsymbol{\omega}, \mathbf{p}) + s^i(\mathbf{p}).\end{aligned} \tag{4}$$

By applying the Funk-Hecke theorem we obtain a harmonic expansion ($N = \infty$) of the term $\alpha(\boldsymbol{\omega}, \mathbf{p})$ as:

$$\alpha(\boldsymbol{\omega}, \mathbf{p}) = \sum_{n=0}^{N} \sum_{m=-n}^{n} (l_{n,m}^{i,c} k_n) h_{n,m}(\mathbf{n}(\mathbf{p})), \tag{5}$$

where $\{h_{n,m}\}$ represents the orthogonal spherical harmonics, and $\{l_{n,m}^{i,c}\}$ and $\{k_n\}$ indicate the expansion coefficients of l_c^i and k with respect to $\{h_{n,m}\}$, respectively. According to [1], most energy in Eq. (5) can be modeled by low-order terms and, therefore, a first- or second-order spherical harmonic approximation can be considered. Particularly, for a distant lighting the 75% of the resulting irradiance is captured for $N = 1$, and the 98% for $N = 2$. Then, higher-order approximations are unnecessary.

As a consequence, the image formation model after including the harmonic expansion is written as:

$$\mathcal{I}_c^i(\mathbf{p}) \approx \rho_c(\mathbf{p})\; \mathbf{l}_c^{i\top} \mathbf{h}[\mathbf{n}](\mathbf{p}) + s^i(\mathbf{p}), \tag{6}$$

where $\mathbf{l}_c^i \in \mathbb{R}^9$ and $\mathbf{h}[\mathbf{n}] \in \mathbb{R}^9$ are the second-order harmonic lighting coefficients and images, respectively, with:

$$\mathbf{h}[\mathbf{n}] = [1, \mathbf{n}_1, \mathbf{n}_2, \mathbf{n}_3, \mathbf{n}_1\mathbf{n}_2, \mathbf{n}_1\mathbf{n}_3, \mathbf{n}_2\mathbf{n}_3, \mathbf{n}_1^2 - \mathbf{n}_2^2, 3\mathbf{n}_3^2 - 1]^\top. \tag{7}$$

Without loss of generality, to consider a first-order approximation (i.e., $N = 1$), we could directly use the first four terms in $\mathbf{h}[\mathbf{n}]$.

III. Variational Photometric Stereo Optimization

We propose to use the image formation model introduced in the previous section to recover non-Lambertian objects in 3D. However, the estimation of normal vectors is a non-linear problem and, as a consequence, we first introduce a model to achieve a linear dependency on depth. This section is devoted to describing the details of our variational approach for uncalibrated specular-aware PS.

A. Measurement Model

We next describe how the process of observing the 3D surface is modeled. Given the 3D coordinates of a surface point in the camera coordinate system $\mathcal{C}$, $\mathbf{x} = [x, y, z]^\top$, and assuming the z-axis aligned with the optical axis of the camera, under a perspective projection the 3D coordinates can be given by:

$$\mathbf{x}(u, v) = z(u, v) \begin{bmatrix} f_u & 0 & u_o \\ 0 & f_v & v_0 \\ 0 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix}, \tag{8}$$

where (f_u, f_v) include the focal length values, (u_o, v_o) the principal point coordinates, and $\mathbf{p} = [u, v, 1]^\top \in \mathcal{M}$ the 2D observation of the point $\mathbf{x}$ in the image plane coded in homogeneous coordinates.

As our goal is to recover the 3D object, we need a way to parametrize the surface normal vector $\mathbf{n}$ at point $\mathbf{x}(u, v)$ by its depth z. To this end, we first compute a vector pointing to the camera as $\bar{\mathbf{n}}(u, v) \approx \partial_u \mathbf{x}(u, v) \times \partial_v \mathbf{x}(u, v)$. According to [4], [13], this normal vector $\bar{\mathbf{n}}[z](u, v)$ is given by:

$$\begin{bmatrix} -\frac{z(u,v)\partial_u z(u,v)}{f_v} \\ -\frac{z(u,v)\partial_v z(u,v)}{f_u} \\ \frac{u-u_0}{f_u}\frac{z(u,v)\partial_u z(u,v)}{f_v} + \frac{v-v_0}{f_v}\frac{z(u,v)\partial_v z(u,v)}{f_u} + \frac{z(u,v)^2}{f_u f_v} \end{bmatrix}, \tag{9}$$

that can be simplified to:

$$\begin{bmatrix} f_u \partial_u z(u, v) \\ f_v \partial_v z(u, v) \\ -(u - u_0)\partial_u z(u, v) - (v - v_0)\partial_v z(u, v) - z(u, v) \end{bmatrix}. \tag{10}$$

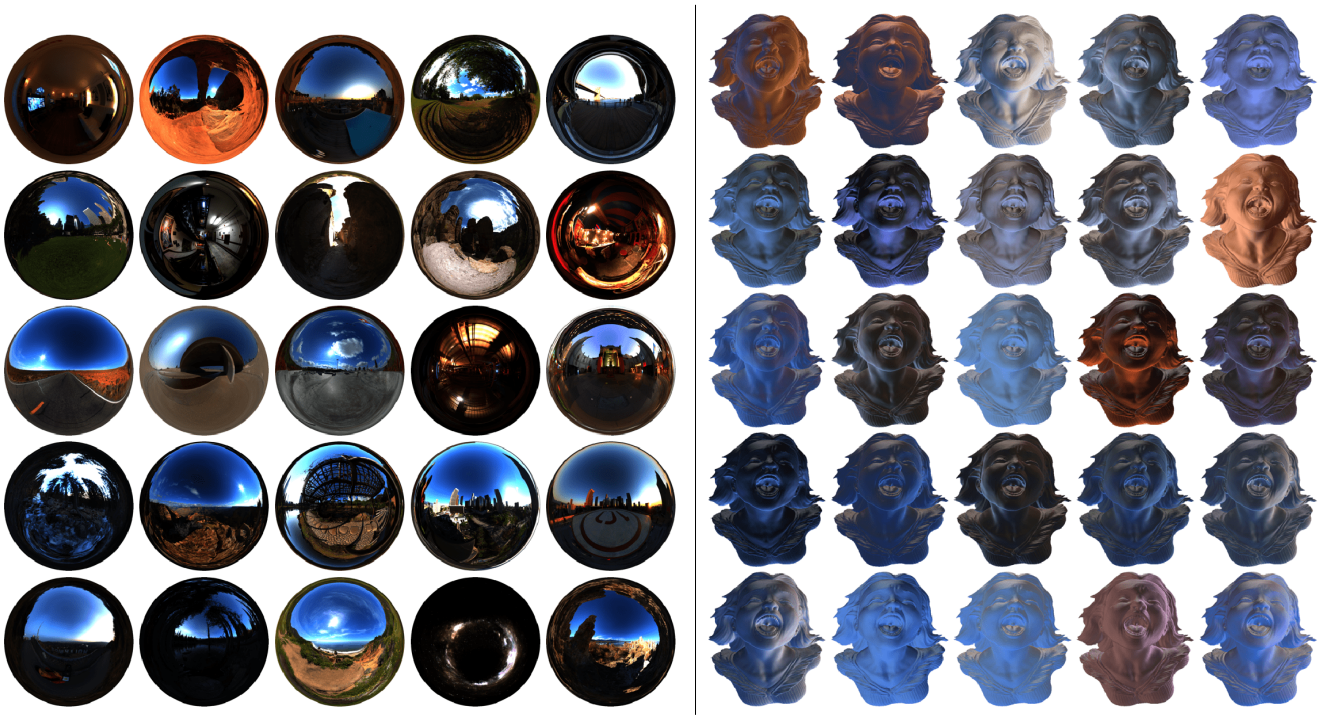

Fig. 1. **Illustration of the synthetic** *Joyful Yell* **dataset**. **Left:** 360-degree spherical environment maps. **Right:** Input images after projecting the *Joyful Yell* mesh with specularity and white albedo. Best viewed in color.

Finally, the unit vector oriented towards the camera is computed as:

$$\mathbf{n}[z](u,v) = \frac{\bar{\mathbf{n}}[z](u,v)}{|\bar{\mathbf{n}}[z](u,v)|}. \tag{11}$$

B. Variational Optimization

Considering our final model in Eq. (6) together with the parametrization in Eq. (11), the model parameters can be simultaneously estimated by minimizing a photometric error of all the observed points over all images by means of the following variational cost function $\mathcal{A}(\{\rho_c\},\{\mathbf{l}_c^i\},\{s^i\},z)$:

$$\begin{aligned}\sum_{i=1}^{I}\sum_{c=1}^{C}\int_{\mathcal{M}} \psi_\lambda\Big(\rho_c(u,v)\,\mathbf{l}_c^{i\,\top}\,\mathbf{h}[\mathbf{n}[z]](u,v) \;+\; s^i(u,v) \\ -\mathcal{I}_c^i(u,v)\Big)\;\mathrm{d}u\;\mathrm{d}v \;\; +\mu\sum_{c=1}^{C}\int_{\mathcal{M}} |\bigtriangledown\rho_c(u,v)|_\gamma\;\mathrm{d}u\;\mathrm{d}v \\ +\mu_s\sum_{i=1}^{I}\int_{\mathcal{M}} |s^i(u,v)|_{\gamma_s}\;\mathrm{d}u\;\mathrm{d}v,\end{aligned} \tag{12}$$

where $\bigtriangledown$ is the spatial gradient operator and $|\cdot|_\gamma$ denotes a Huber norm. In the data term we use a Cauchy's M-estimator defined by $\psi_\lambda(q) = \lambda^2\log(1+\frac{q^2}{\lambda^2})$, where λ is a scaling coefficient. Unfortunately, the previous problem is non-convex and highly non-linear.

To better condition it, we add two regularization priors to improve the joint solution. The first one consists of a Huber total variation term on each albedo map and it is used to enforce smoothness on the albedo maps $\{\rho_c\}$. To prevent a full explanation of the image as a specularity, we also penalize its use by means of a second Huber-based regularizator. Both regularizators are balanced with the data term by using the weights $\{\mu,\mu_s\} > 0$ that are determined empirically. The Huber loss can be defined as:

$$|q|_\gamma = \begin{cases} |q|^2/(2\gamma) & \text{if } |q| \le \gamma \\ |q| - \gamma/2 & \text{if } |q| > \gamma \end{cases}, \tag{13}$$

where γ is a fixed coefficient.

It is worth noting that thanks to our formulation in Eq. (12), we can directly solve the uncalibrated PS problem in terms of z instead of recovering a set of potentially non-integrable normal vectors. Moreover, that means an improvement in terms of computational efficiency as additional post-processing steps are unnecessary.

C. Implementation and Initialization

To numerically solve the problem in Eq. (12), the domain $\mathcal{M}$ is replaced by the number of pixels P, and employing a forward difference stencil to discretize the spatial gradient we can finally obtain discretized vectors to encode our model parameters. Then, the optimization problem can be efficiently solved by means of a lagged block coordinate descent algorithm (see supplementary for more details). Basically, the full problem is tackled by means of partial subproblems independently resolved via least squares strategies. As our variational formulation is non-convex, it is important not to initialize their values at random, and for that reason, we follow a strategy as in [4] to initialize albedos and lighting. In addition to that, we impose null specularity for initialization.

Regarding the depth initialization, we use image silhouette to recover a balloon-like surface by solving a constrained minimal surface problem [14], [20], [21]. Particularly, the solution is constrained by a volume value κ that is set a priori. Fortunately, as in real-world applications the distance from the camera to the object shape is within reasonable bounds, the relation between shape area and shape volume is always similar, simplifying the search for a volume value κ. We will consider that in the experimental section.

IV. Experimental Evaluation

In this section we show experimental results on both synthetic and real image collections providing both qualitative and quantitative comparison with respect to state-of-the-art-solutions. For quantitative evaluation, we consider the mean angular error between ground truth $\mathbf{n}^{GT}[z]$ and estimated $\mathbf{n}[z]$ normals defined as:

$$\text{MAE} = \frac{1}{P}\sum_{p=1}^{P} \tan^{-1}\Big(\frac{|\mathbf{n}^{GT}[z] \times \mathbf{n}[z]|}{\mathbf{n}^{GT}[z] \cdot \mathbf{n}[z]}\Big), \qquad (14)$$

where $\times$ and $\cdot$ denote cross and dot products, respectively. In all experiments, we set $\gamma = \gamma_s = 0.1$ and $\lambda = 0.15$. The rest of coefficients will be considered later.

A. Synthetic Data

We propose the use of four synthetic shapes with different light conditions for evaluation. Particularly, we consider the challenging shapes *Joyful Yell* provided by [27], and *Armadillo*, *Lucy*, and *ThaiStatue* provided by [10]. To generate the datasets, we employ 25 environment maps l^i from [6] with a white albedo ($\rho_c(\mathbf{p}) = 1$). To render synthetic input images, we use Eq. (1) in combination with a specularity mask ($s^i(\mathbf{p}) \neq 0$) per image. All the environment maps we use are shown in Fig. 1-left, as well as the impact of every incident lighting in combination with specularity over the *Joyful Yell* shape in Fig. 1-right.

First of all, we analyze the effect of the initialization in our approach. It is worth recalling that our variational formulation is non-convex and, as a consequence, it is important to consider initialization of certain parameters. To do that, we have tried three alternatives for the specular term: full specularity ($s^i(\mathbf{p}) = 1$), null specularity ($s^i(\mathbf{p}) = 0$), and the provided by [18]. The last one is based on the fact that the intensity ratios for diffuse pixels are independent of the shape geometry. Figure 2 shows visually the effect of those initializations in comparison with the specular ground truth as well as the estimated 3D shape. As it can be seen, when full specularity is used that component absorbs too much information, which causes a very poor estimation of the depth. For null specularity, both the estimated specular component and the 3D shape are very competitive. Finally, when we use [18] for initialization following the implementation in [17], the 3D shape we obtain is slightly worse than the one obtained by ($s^i(\mathbf{p}) = 0$), but it converges faster to the solution. Thus, the estimation obtained by the last initialization could provide the best trade-off between accuracy and efficiency. However,

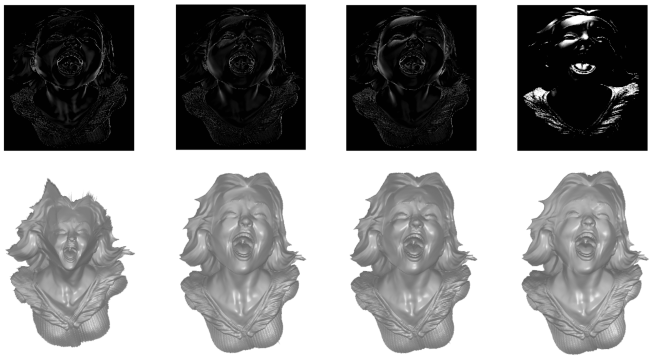

Fig. 2. **Effect of specular initialization.** In all cases, the figure displays the case of full specularity at initialization, null specularity at initialization, initialization by [17], and ground truth for the *Joyful Yell* shape. **Top:** Specular maps **Bottom:** Depth estimation.

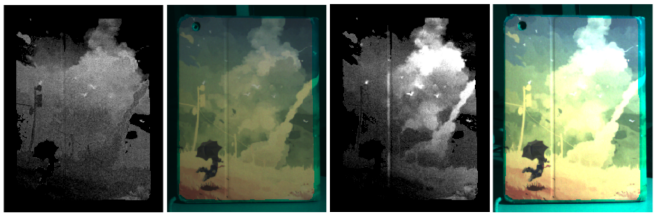

Fig. 3. **Specular initialization in white areas.** Comparison in objects without (on the left) and with specularity (on the right). Odd columns show the initial specular component by [17] and, in the pair ones the albedo-aware ground truth images.

the method fails for images with negligible specular lighting by considering white pixels as points with specularity. This effect is shown in Fig. 3 where some white parts of the object are initialized with a high specularity. For that reason, we will finally assume null specularity for initialization, as our goal is to achieve the most accurate method as possible while robustness is not compromised.

We next evaluate the effect of depth initialization in section III-C. To discover the optimal volume values κ for each dataset, we consider the range of values $[10^0, 10^2]$. As it can be seen in Fig. 4-left, the solution is stable for a part of the interval. Without loss of generality, for these synthetic datasets we chose the values which provided the most accurate shapes, and fixed them for the rest of the experiments. The optimal values were 44, 12.25, 8.25 and 8 for *Joyful Yell*, *Armadillo*, *Lucy*, and *ThaiStatue* datasets, respectively.

Now, we evaluate the effect of the regularization weights $\{\mu, \mu_s\}$ in Eq. (12). To this end, we consider for both weights a large range of values for $[10^{-7}, 10^{-4}]$. The MAE error for every dataset is represented as a function of those μ_s and μ values in Fig. 4-right and Fig. 4-middle, respectively. Considering all the datasets, we fix for the rest of experiments $\mu_s = 2 \cdot 10^{-6}$ and $\mu = 3 \cdot 10^{-6}$ despite not being the optimal values for every dataset independently.

Finally, we provide quantitative evaluation and comparison with respect to UPS [4], the most competitive technique in state of the art for our approach according to table I. This method provided better performance compared to [12], [15], [19] in [4]. The parameters of this method were set in

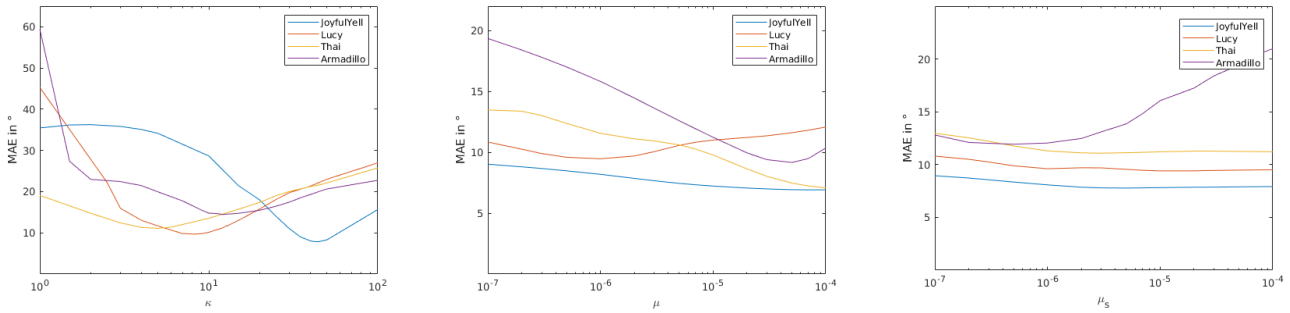

Fig. 4. **MAE errors to measure the precision on the estimated depth as a function of tuning parameters for synthetic datasets. Left:** Initial volume evaluation. **Middle:** Regularization on albedos by μ. **Right:** Regularization on specularities by μ_s.

Meth. \ Dataset	Joyful Yell	Armadillo	Lucy	ThaiStatue	Average
Ours	7.66	13.63	10.05	10.93	10.66
UPS [4]	13.44	24.83	14.43	23.74	19.11
Relative incre.	1.754	1.822	1.436	2.172	1.793

TABLE II

RECONSTRUCTION QUANTITATIVE COMPARISON. THE TABLE REPORTS THE MAE RESULTS IN DEGREES FOR UPS [4] AND OUR ALGORITHM. RELATIVE INCREMENT IS COMPUTED WITH RESPECT TO OUR METHOD, THE MOST ACCURATE SOLUTION.

accordance with the original paper. Our results are summarized in Table II. As it can be observed, the provided method can achieve a MAE error of 10.66 degrees on average, a more accurate solution than the 19.11 degrees obtained by UPS [4]. In other words, our approach reduces the error a 79.3%. As expected, UPS [4] cannot handle scenarios with specularities as good as our method does. A qualitative comparison with the competing technique as well as with the ground truth can be seen in Fig. 5. Both methods provide physically coherent estimations, but the differences of the 3D reconstructions are noticeable in the *Armadillo* and *ThaiStatue* datasets as more details are correctly acquired by our approach.

B. Real Data

In this section, we provide qualitative evaluation and comparison on four different datasets. Particularly, we consider the real-world shapes *Ovenmitt*, *Tablet*, *Face*, and *Vase* [3]. This set of datasets offers a large variety of complex geometries (nearly planar, smooth and wrinkled shapes) and albedos and was captured under daylight and a freely moving LED. Depth initialization was manually set by exploiting the relation between shape area and shape volume, as it was commented above. The results are displayed in Fig. 6. As it is showed, our approach visually obtains better 3D representations, especially for *Ovenmitt*, *Tablet*, and *Vase* datasets, by recovering more spatial details as well as achieving a global consistency. In addition to that, reflectance seems to be slightly more accurate, as it can be seen in the *Vase* dataset. Similar results with respect to UPS [4] are obtained in the *Face* dataset, as the specular component is low. In this line, we consider that our method outperforms UPS [4] for the *Vase* dataset due to the specular component in this set of images being bigger and, as a consequence, our method can handle it better and produce

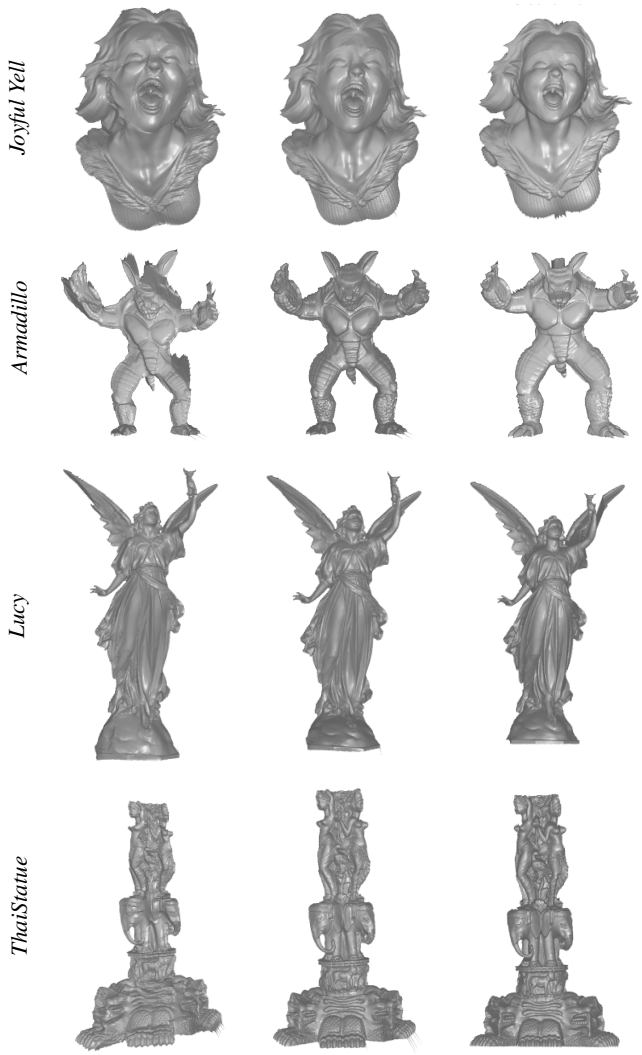

Fig. 5. **Qualitative comparison on synthetic datasets.** Comparing our estimation with respect to UPS [4] and the ground truth. From left to right it is displayed the UPS [4] estimation, our solution, and ground truth.

more accurate joint solutions. This can be easily observed in Fig. 7, where we show our specular estimations for one particular image of several datasets. Fortunately, our approach can capture properly the specular component, making more robust the joint estimation of depth, albedo, and specularity.

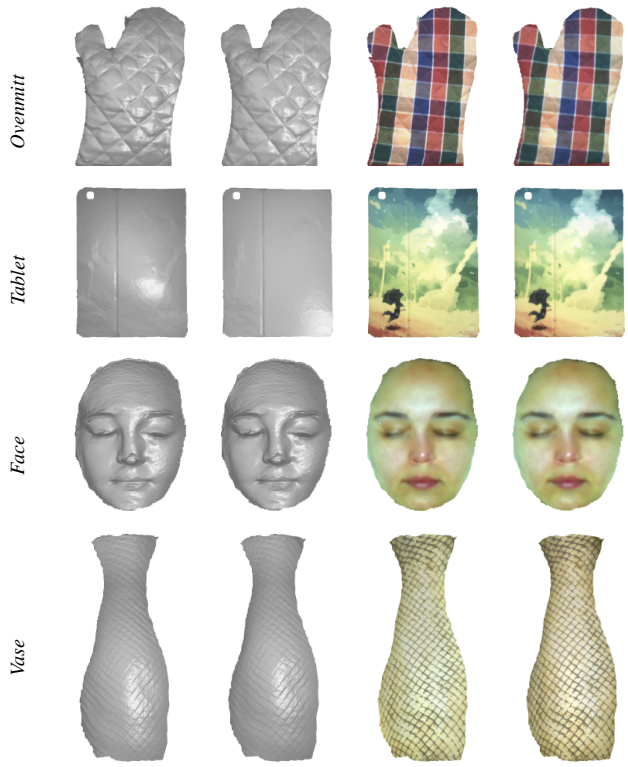

Fig. 6. **Qualitative comparison on real datasets.** Comparing our estimation with respect to UPS [4]. First and Third columns: depth and albedo by UPS [4]. Second and Fourth columns: depth and albedo by ours.

Meth. \ Dataset	Vase	Tablet	Ovenmitt	Face2	Average
Ours	705.32	884.62	765.34	295.88	662.79
UPS [4]	627.45	783.24	667.16	257.62	583.86
Relative incre.	1.124	1.129	1.147	1.148	1.135

TABLE III
COMPUTATION TIME QUANTITATIVE COMPARISON. THE TABLE REPORTS THE COMPUTATION TIME RESULTS IN SECONDS FOR UPS [4] AND OUR ALGORITHM. RELATIVE INCREMENT IS COMPUTED WITH RESPECT TO UPS [4], THE MOST EFFICIENT SOLUTION.

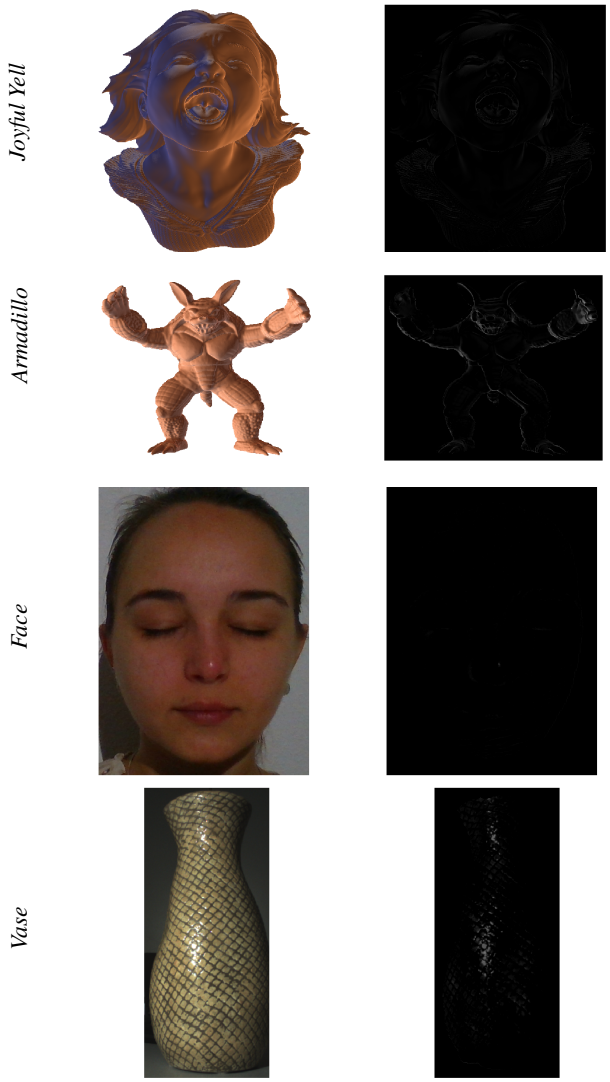

Fig. 7. **Specular Recovery.** Some input images for synthetic and real images (left columns) together with the specular estimations (right column) obtained by our method.

For instance, our method can capture some specularities that are not clearly a diffuse component, as it can be seen for the nose in the *Face* dataset (see Fig. 7, third row). Regarding computational complexity, as our method can provide more complete estimations, the computation time is slightly larger than the provided by UPS [4]. A summary of those results (in non-optimized Matlab code) on a commodity laptop Intel Core i7-8700 3.20GHz CPU are reported in Table III for real datasets. Despite increasing the computational time a 13.5%, the trade-off between accuracy and computation time is very competitive for our approach as it can accurately solve the problem with a small increase in computation.

V. CONCLUSION

The uncalibrated PS problem under general lighting has been approached by a variational and unified method. The proposed algorithm can jointly retrieve reflectance, 3D reconstruction, lighting and specularities of the object, all of them, from a set of RGB images and without assuming any training data at all. To this end, we have presented a physical-aware image formation model that, in combination with a perspective projection one and under spherical harmonics lighting, gives a fully interpretable algorithm. Thanks to our formulation, we can handle a large variety of complex geometries and illumination conditions without needing any knowledge prior. We have experimentally evaluated our approach both on synthetic and real datasets. While our approach provides a full and interpretable model capable of generating competitive joint 3D and lighting reconstructions with respect to competing techniques, the computational cost slightly increases. An interesting avenue for future research would be to validate our formulation for articulated objects where strong shadows could appear due to self-occlusions.

Acknowledgment: This work has been partially supported by the Spanish Ministry of Science and Innovation under project MoHuCo PID2020-120049RB. The authors wish to thank B. Haefner for fruitful discussions.

REFERENCES

[1] R. Basri, D. Jacobs, and I. Kemelmacher, "Photometric stereo with general, unknown lighting," *International Journal of Computer Vision*, vol. 72, no. 5, pp. 239–257, 2007.

[2] M. Chandraker, J. Bai, and R. Ramamoorthi, "On differential photometric reconstruction for unknown, isotropic BRDFs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2941–2955, 2013.

[3] B. Haefner, S. Peng, A. Verma, Y. Queau, and D. Cremers, "Photometric depth super-resolution," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 10, pp. 2453–2464, 2019.

[4] B. Haefner, Z. Ye, M. Gao, T. Wu, Y. Quéau, and D. Cremers, "Variational uncalibrated photometric stereo under general lightings," in *IEEE International Conference on Computer Vision*, 2019.

[5] 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.

[6] HDRLabs, "sIBL archive," URL: http://www.hdrlabs.com/sibl/archive.html.

[7] Y. Hold-Geoffroy, P. Gotardo, and J. Lalonde, "Deep photometric stereo on a sunny day," in *Arxiv preprint 1803.10850*, 2018.

[8] Y. Hold-Geoffroy, J. Zhang, P. Gotardo, and J. Lalonde, "What is a good day for outdoor photometric stereo?" in *International Conference on Computational Photography*, 2015.

[9] S. Ikehata, "CNN-PS: CNN-based photometric stereo for general non-convex surfaces," in *European Conference on Computer Vision*, 2018.

[10] M. Levoy, J. Gerth, B. Curless, and K. Pull, "The stanford 3D scanning repository," 2005.

[11] J. Li, A. Robles-Kelly, S. You, and Y. Matsushita, "Learning to minify photometric stereo," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[12] Z. Mo, B. Shi, F. Lu, S. Yeung, and Y. Matsushitae, "Uncalibrated photometric stereo under natural illumination," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[13] G. Munda, J. Balzer, S. Soatto, and T. Pock, "Efficient minimal-surface regularization of perspective depth maps in variational stereo," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[14] M. R. Oswald, E. Toeppe, and D. Cremers, "Fast and globally optimal single view reconstruction of curved objects," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[15] S. Peng, B. Haefner, Y. Queau, and D. Cremers, "Depth super-resolution meets uncalibrated photometric stereo," in *IEEE International Conference on Computer Vision Workshops*, 2017.

[16] Y. Queau, T. Wu, F. Lauze, J. Durou, and D. Cremers, "A non-convex variational approach to photometric stereo under inaccurate lighting," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[17] V. Ramos, "SIHR: A matlab/GNU octave toolbox for single image highlight removal," *Journal of Open Source Software*, vol. 5, no. 45, p. 1822, 2020.

[18] H. L. Shen and Z. H. Zheng, "Real-time highlight removal using intensity ratio," *Applied Optics*, vol. 52, pp. 4483–4493, 2013.

[19] B. Shi, K. Inose, Y. Matsushita, P. Tan, S. Yeung, and K. Ikeuchi, "Photometric stereo using internet images," in *3D Vision*, 2014.

[20] E. Toeppe, M. R. Oswald, D. Cremers, and C. Rother, "Silhouette-based variational methods for single view reconstruction," in *Video Processing and Computational Video*, 2010.

[21] S. Vicente and L. Agapito, "Balloon shapes: reconstructing and deforming objects with volume from images," in *3D Vision*, 2013.

[22] X. Wang, Z. Jian, and M. Ren, "Non-lambertian photometric stereo network based on inverse reflectance model with collocated light," *IEEE Transactions on Image Processing*, vol. 29, pp. 6032–6042, 2020.

[23] R. Woodham, "Photometric method for determining surface orientation from multiple images," *Optical Engineering*, vol. 19, no. 1, p. 191139, 1980.

[24] L. Wu, A. Ganesh, B. Shi, Y. Matsushita, Y. Wang, and Y. Ma, "Robust photometric stereo via low-rank matrix completion and recovery," in *Asian Conference on Computer Vision*, 2010.

[25] T. Wu and C. Tang, "Photometric stereo via expectation maximization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 3, pp. 546–560, 2007.

[26] Z. Yao, K. Li, Y. Fu, H. Hu, and B. Shi, "GPS-net: Graph-based photometric stereo network," in *Conference on Neural Information Processing Systems*, 2020.

[27] T. J. Yell, URL: http://www.thingiverse.com/thing:897412.