DETAIL-AWARE UNCALIBRATED PHOTOMETRIC STEREO

Antonio Agudo

Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Spain

ABSTRACT

Photometric stereo is the problem of jointly inferring the 3D reconstruction, reflectance, lighting and specularities of an object from a set of visual signals. Recently, some variational, uncalibrated, unsupervised and unified formulations have provided robust solutions to the problem while reducing the prior knowledge about the shape geometry or the lighting conditions. Unfortunately, these approaches cannot still produce solutions with an ample variety of details in the shape. That is mainly due to the non-convex and non-linear nature of the problem which requires the best initialization as possible. In this context, we propose a fully interpretable formulation that combines a physically-aware image formation model under perspective projection with a minimal detail-aware initialization and that it can handle general lighting. As a result, our formulation can consider multiple scenarios composed of unknown complex geometries and lighting patterns. Experimental results on challenging synthetic and real datasets show the effectiveness of our approach to capture more fine details, outperforming state-of-the-art techniques in terms of 3D reconstruction.

Index Terms— Uncalibrated Photometric Stereo, Minimal Surfaces, Unsupervised Vision, Specular Materials.

1. INTRODUCTION

Simultaneously recovering the 3D reconstruction of a rigid object, its reflectance, lighting and specularities from multiple visual signals taken at the same viewpoint but under different illumination conditions is coined in literature as Photometric Stereo (PS) [1, 2, 3, 4]. Firstly, the most standard way to solve the problem was to invert a physically-aware image formation model by assuming a certain control of lighting. In spite of providing robust and accurate solutions, these approaches drastically reduced its applicability to laboratory setups where an exhaustive calibration of lighting was mandatory. In the counterpart, uncalibrated formulations were presented to avoid the strong assumptions of the calibrated ones, obtaining an ill-posed problem as the underlying normal map to encode the shape is recovered up to a linear ambiguity [5]. Later, the ambiguities were relaxed by means of the use of variational formulations where the 3D reconstruction was encoded as a depth map rather than using normal vectors [6, 7], but just for a single source of illumination. Handling general lighting in an uncalibrated fashion [8, 9, 10, 11, 12] is a very important problem in signal processing with potentially many real-world applications in art conservation, agriculture, biology, the movie industry and, motion capture of humans and animals, to name just a few. These works were proposed to cope with Lambertian [11] and non-Lambertian [12] materials, but the quality of their solutions were strongly influenced by the initialization step [13, 14, 15]. In practice, the final 3D estimation does not include many details as this step represents a key factor in the full algorithm.

Secondly, PS has been also solved by deep-learning approaches [4, 16, 17, 18, 19]. Basically, this family of methods exploits end-to-end learning architectures where some ground-truth parameters are used for supervision in training, obtaining implicit relations of the parameters to be estimated. Despite obtaining promising results, their lack of physical interpretability prevents them from knowing the real interactions between lighting, surface 3D geometry and specularities. Nonetheless, the Achilles' heel of these proposals is the lack of large amounts of data for training, that should include 3D geometries of complex objects, the knowledge of general lighting, or the specularities. Unfortunately, the acquisition of this set of ground truths, even in a controlled lab setting, is very laborious and expensive or unattainable for certain types of materials. As a consequence, these methods tend to include previous assumptions in the formulation, such as the light direction or the intensity at every instant, i.e., a type of calibration.

We overcome most of the limitations of current methods by proposing a variational, unsupervised, unified, and uncalibrated PS algorithm that can work under general lighting and it is available for non-specific objects and materials. Our approach does not need any extra sensor, training data or ground truth of any kind, it is fully interpretable and can run in a commodity laptop in an efficient way. Instead, we propose a convex formulation to initialize the 3D geometry from a single visual signal that can recover a finer level of details and it represents a key factor in the final estimation of our algorithm. As this algorithm is not based on learning, the formulation may process any object indistinctly, even including those with thin areas.

This work has been supported by the project MoHuCo PID2020-120049RB-I00 funded by MCIN/ AEI /10.13039/501100011033.

2. PHYSICALLY-AWARE PHOTOMETRIC STEREO

Let $\{\mathcal{I}_c^i \subset \mathbb{R}^2\}$ be a set of $i = \{1, \dots, I\}$ visual signals with different illumination conditions and with $c = \{1, \dots, C\}$ color channels where a rigid object to be reconstructed appears. For that object, we also define $\mathcal{S} \subset \mathcal{I}_c^i$ as its shape segmentation in the signal set, and by means of $\mathcal{B} \subset \mathcal{S}$ its silhouette boundary, i.e., $\mathcal{S}$ contains P pixels inside the silhouette of the object shape and $\mathcal{B}$ only its boundary. Inspired by a Phong reflection model [20], the light at a p-th pixel can be modeled as the sum of two additive terms: a viewpoint-independent diffuse and a view-dependent specular. With that in mind, the surface reflectance for all P pixel points $\mathbf{p} = [u, v]^\top \in \mathcal{S}$ can be modeled by collecting elementary luminance contributions arising from all the incident lighting directions $\boldsymbol{\omega}$ as:

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

where $\mathbb{S}^2$ indicate the unit sphere in $\mathbb{R}^3$, $\rho_c(\mathbf{p}) \in \mathbb{R}^+$ and $l_c^i(\boldsymbol{\omega})$ represent the color-wise albedos and intensity of the incident lights, respectively, $\mathbf{n}(\mathbf{p})$ the unit-length surface normal at the surface point conjugate to p-th pixel and, $s^i(\mathbf{p})$ for the specular reflection. The object irradiance or shading component is coded by the max operator.

With these ingredients, the PS problem [1, 3, 6] in an uncalibrated fashion consists in retrieving the 3D shape of the object (via its normal vectors $\mathbf{n}(\mathbf{p})$) together with the quantities $\{\rho_c\}$, $\{l_c^i\}$ and $\{s^i\}$, all of them, from the signal set $\{\mathcal{I}_c^i\}$. To do the problem tractable, many works in literature encode the irradiance map by spherical harmonics of general lighting [21], considering first- or second-order approximations. Image formation model in Eq. (1) could be now written as:

$$\mathcal{I}_c^i(\mathbf{p}) \approx \rho_c(\mathbf{p})\, {\mathbf{l}_c^i}^\top \mathbf{h}[\mathbf{n}](\mathbf{p}) + s^i(\mathbf{p}), \quad (2)$$

where $\mathbf{l}_c^i \in \mathbb{R}^o$ and $\mathbf{h}[\mathbf{n}] \in \mathbb{R}^o$ with $o \in \{4, 9\}$ are the first- or second-order harmonic lighting coefficients and images, respectively. To avoid the non-linear problem of estimating normal vectors, every surface normal vector $\mathbf{n}[z]$ is parametrized by its depth [11, 12] under a perspective projection.

Energy Formulation. The model parameters can be estimated by minimizing an image render error of all the observed points over all visual signals. To this end, a residual function $g_{c,p}^i$ is defined to represent the error between the predicted intensity and the real one at the p-th pixel as:

$$g_{c,p}^i(\beta_p, \rho_{c,p}, \mathbf{l}_c^i, s_p^i, z) = \rho_{c,p}\, {\mathbf{l}_c^i}^\top \mathbf{h}_p[\bar{\mathbf{n}}_p[z]/\beta_p] + s_p^i - \mathcal{I}_{c,p}^i,$$

where $\beta_p = |\bar{\mathbf{n}}_p[z]|\ \forall p \in \{1, ..., P\}$ and $\mathbf{n}_p[z] = \bar{\mathbf{n}}_p[z]/\beta_p$, with $\bar{\mathbf{n}}_p[z]$ a linear parametrization on depth. In addition to the data term, two regularization priors to enforce smoothness on the albedo and specular maps are considered. Then, the total cost function $\mathcal{A}(\boldsymbol{\beta}, \{\rho_c\}, \{\mathbf{l}_c^i\}, \{s^i\}, z)$ can be written as:

$$\sum_{i=1}^{I}\sum_{c=1}^{C}\sum_{p=1}^{P} \psi_\lambda(g_{c,p}^i(\beta_p, \rho_{c,p}, \mathbf{l}_c^i, s_p^i, z)) + \mu \sum_{c=1}^{C}\sum_{p=1}^{P} |(\bigtriangledown \rho_c)_p|_\gamma + \ \mu_s \sum_{i=1}^{I}\sum_{p=1}^{P} |(s^i)_p|_{\gamma_s}\,, \quad (3)$$

where $\bigtriangledown$ is the spatial gradient operator, $|\cdot|_\gamma$ denotes a Huber norm and $\psi_\lambda(q) = \lambda^2 \log(1 + \frac{q^2}{\lambda^2})$ a Cauchy's M-estimator where λ is a scaling coefficient.

Unfortunately, the previous problem is non-convex and highly non-linear and, as a consequence, a proper initialization represents a key factor to obtain accurate results as we will see below. In any case, a lagged block coordinate descent algorithm is used to minimize $\mathcal{A}$ in Eq. (3).

Minimal Detailed Surface. To initialize depth, we propose to solve a minimal problem to infer detailed surfaces from a single image, i.e., estimating a depth value for every p-th pixel in $\mathcal{S}$. As input we consider a grey average image as $\bar{\mathcal{I}} = \text{mean}(\mathcal{I}_c^i)$. To that end, we propose to minimize a new energy $\mathcal{D}(z)$ function composed of a data term [14, 15], a shape-detail regularizator [22] and a volume one as:

$$\sum_{p=1}^{P} \sqrt{1 + |(\nabla z)_p|^2} + \theta(z_p - w_p)^2 + \upsilon \sum_{p=1}^{P} z_p - V, \quad (4)$$

where θ and υ are weight coefficients and w_p is a function to regularize the solution. V indicates the volume of the object and it is useful to scale the monocular reconstruction. Note that this formulation never considers any depth real value to constrain the solution and, therefore, our global formulation estimates the 3D just from 2D visual signals.

Following [13, 14, 15], the thickness of the object increases as we move inward from its silhouette boundary B, especially in nature [22]. However, the use of a basic data term to solve the minimal problem cannot recover a wide variety of details as well as thin regions. To avoid that, we first consider the distance $d(p, \partial B)$ to the boundary B for any interior point $p \in S$ as $d(p, \partial B) = \min_{b \in \partial B} \|p - b\|$ (see Fig. 1-second column). Moreover, a detail map $e(\bar{\mathcal{I}})$ by exploiting image information as $e(\bar{\mathcal{I}}) = \zeta \frac{(|\nabla \bar{\mathcal{I}}| - \min(|\nabla \bar{\mathcal{I}}|))}{\max(|\nabla \bar{\mathcal{I}}|) - \min(|\nabla \bar{\mathcal{I}}|)}$ is considered (an example is displayed in Fig. 1-third column), where ζ is a weight coefficient, and $e(\bar{\mathcal{I}}_m) = 0$ for any point $m \notin S$. The previous terms are now combined (see Fig. 1-fourth column) to define the function w, that for the p-th pixel location can be written as:

$$w_p = \min\{\phi, \eta + \kappa\, d(p, \partial B) + e(\bar{\mathcal{I}}_p)\}, \quad (5)$$

where $\{\phi, \eta, \kappa\}$ are weights to encode the type of prior. Particularly, ϕ limits the level of extrusion of the object and it can be tuned as $\phi = \alpha \max(d(p, \partial B))$, with $\alpha \in [0, 1]$. η is to guarantee a minimum of extrusion in those points close to the

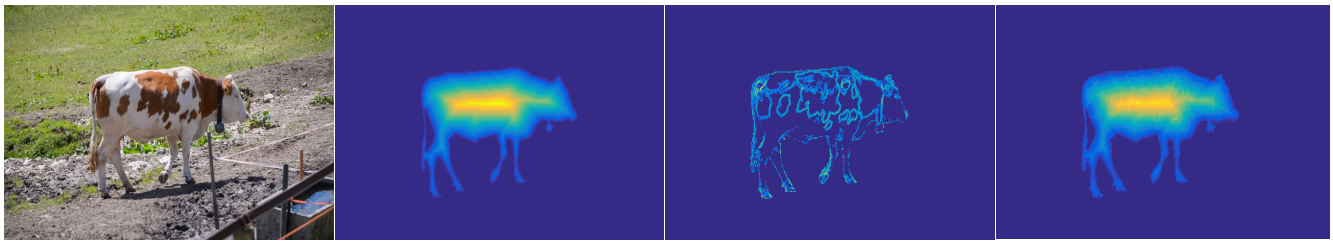

Fig. 1. **Shape prior interpretation**. From left to right, it is displayed a color input visual signal, the distance function $d(\bar{\mathcal{I}}, \partial B)$, the detail map $e(\bar{\mathcal{I}})$, and the overall $w(\bar{\mathcal{I}})$ function to regularize the shape. As it can be seen, every component acts over the final constraint, providing an accurate regularizer to extrude shape surfaces. Best viewed in color.

boundary. Finally, Dirichlet's boundary conditions are also enforced as $z_p = 0, \ \ \forall p \in \mathcal{B}$. As the problem $\mathcal{D}(z)$ in Eq. (4) is convex, an iterative gradient descent method is employed.

3. EXPERIMENTAL EVALUATION

We now present our experimental results on both synthetic and real datasets, providing quantitative evaluation and comparison with respect to competing techniques, as well as a qualitative one. For quantitative evaluation we compute the mean angular error between the estimated $\mathbf{n}[z]$ and ground-truth $\mathbf{n}^{GT}[z]$ normal vectors by means of MAE $= \frac{1}{P}\sum_{p=1}^{P} \tan^{-1}\left(\frac{|\mathbf{n}^{GT}[z]\times\mathbf{n}[z]|}{\mathbf{n}^{GT}[z]\cdot\mathbf{n}[z]}\right)$, where $\times$ and $\cdot$ indicate cross and dot products, respectively. In our experiments, to minimize $\mathcal{A}(\cdot)$ we set $\gamma = \gamma_s = 0.1$, $\mu = 3 \cdot 10^{-5}$, $\mu_s = 2 \cdot 10^{-6}$ and $\lambda = 0.15$; and to minimize $\mathcal{D}(\cdot)$ the values $\theta = 0.1$, $\upsilon = \kappa = 1$, $\eta = 0$, and $\alpha = 0.9$. In our algorithm, we use first-order spherical harmonics in the first eight iterations [11] and then change to second-order ones. Moreover, we always initialize the specularities by null maps.

Synthetic datasets. Four challenging synthetic object shapes with different light conditions are considered: *Joyful Yell* provided by [23]; *Armadillo*, *Lucy*, and *ThaiStatue* provided by [24]. In order to generate the datasets, we follow [12], employing 25 environment maps l^i from [25] with a white albedo ($\rho_c(\mathbf{p}) = 1$) and a specularity mask ($s^i(\mathbf{p}) \neq 0$). The final visual signals we use for evaluation are produced by applying Eq. (1) (some instances are provided in the left column of Fig. 2).

We first use these datasets to evaluate how our full algorithm works, but before proceeding the depth initialization needs to be considered. To this end, we experiment with an interval of V values $[1, 100]$. Despite obtaining stable solutions within, we decide to provide the optimal values for comparison purposes. Particularly, the optimal values were 32, 4, 5, and 3.8 for *Joyful Yell*, *ThaiStatue*, *Armadillo* and *Lucy*, respectively. Without loss of generality, as the distance from object to the camera is within reasonable bounds, the relation between shape area and volume is always quite similar, simplifying the volume value V selection. Regarding the detail-aware minimal initialization, we set $\zeta = 10$ to enforce a natural level of details, avoiding solutions with no details, i.e., $\zeta = 0$, or other types of over-estimations that could occur for higher values. Our results for these values are displayed in Fig. 2. As it can be seen, our solution is physically plausible and seems very close to the ground truth in terms of 3D reconstruction (see right part in the figure). Moreover, we also provide a quantitative evaluation with respect to UPS [11] and SAUPS [12], the most accurate approaches in state of the art. The parameters of these methods were set in accordance with the original papers. Our results are reported in Table 1, obtaining a MAE error of 9.32 on average. It is worth noting that our approach outperforms the competing approaches

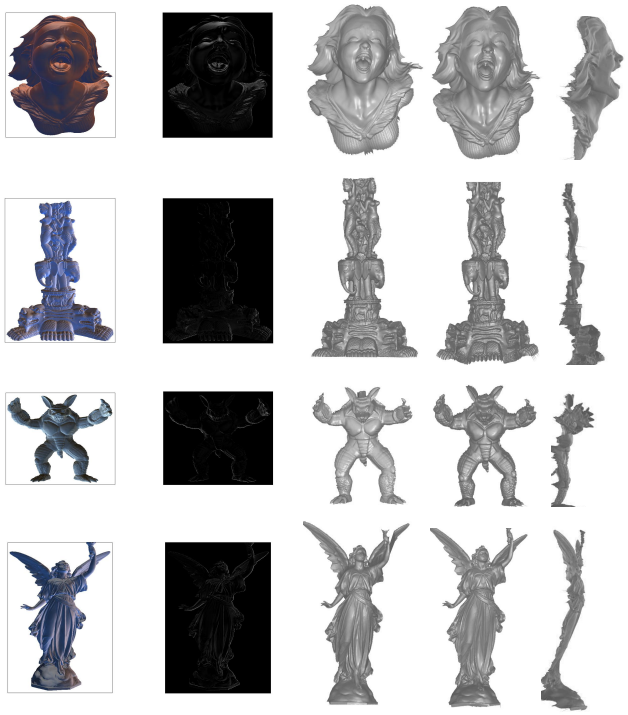

Fig. 2. **Qualitative evaluation on synthetic datasets**. From top to bottom: *Joyful Yell*, *ThaiStatue*, *Armadillo*, and *Lucy*. The same information is shown for the all cases. From left to right, it is displayed an arbitrary i-th input image, the i-th estimated specular map, the ground truth 3D shape, our 3D estimation and an alternative point of view of that reconstruction.

Meth. \ Dataset	Joyful Yell	Armadillo	Lucy	ThaiStatue	Average
UPS [11]	13.44	24.83	14.43	23.74	19.11
SAUPS [12]	7.66	13.63	10.05	10.93	10.66
Ours	**7.28**	**9.61**	**9.86**	**10.55**	9.32
Relative error w.r.t. [11]/ [12]	1.85/ 1.05	2.58/ 1.42	1.46/ 1.02	2.25/ 1.04	2.05/ 1.13

Table 1. **3D reconstruction evaluation and comparison.** The table reports the MAE results in degrees for UPS [11], SAUPS [12] and, our algorithm. Relative errors are computed with respect to our solution, the most accurate algorithm on average.

SAUPS [12] and UPS [11] by large margins between the 13% and 205% on average, respectively. Our method obtains the best performance for the *Joyful Yell* dataset, that includes a wide variety of complex areas with many details. In any case, as it can be observed in the figure, our approach can capture most of details in all the datasets. In addition to that, second column in the figure also shows some instances of specular estimation our algorithm can infer.

Real datasets. We now present a qualitative evaluation on four real-world visual collections. In particular, the shapes to be captured represent an ample variety of natural geometries, including smooth, nearly planar, and areas with a high level of details as well as different examples of natural albedos. The visual collections are *Ovenmitt*, *Tablet*, *Face*, and *Vase* [26], and they were captured under daylight and a freely moving LED.

For depth initialization, as no prior knowledge is available, we exploit the relation between shape area and volume to set a volume as it was commented above. Our joint estimations are shown in Fig. 3. First, we analyze the 3D reconstruction our algorithm can infer. As it can be seen, in all cases the estimation is physically possible, globally consistent and compatible with the input images, being most of the details captured (see fourth and fifth columns in the figure). However, we can still observe some non-smooth regions in the *Face* scenario, or a bit of over-deformations in the *Vase* collection that could be avoided by including a shape regularization to our energy in Eq. (3). It is worth mentioning that our method can even recover naturally the *Tablet* scenario, where the 3D shape is a challenging nearly planar shape. Besides that, in the third column of the figure are displayed the estimated albedos by our algorithm. Again, these solutions seem to be accurate in accordance with the input image (see first column in the same figure). Regarding specularities, our algorithm captures a bigger contribution in the *Vase* dataset, and some challenging local specularities in the *Ovenmitt*, *Tablet*, and *Face*. Particularly noteworthy is the local estimation at the tip of the nose in the *Face* dataset. These results can be seen in the second column of Fig. 3, where some instances of recovered specularities are displayed. On balance, as our algorithm can capture accurately both specularies and reflectances, the corresponding 3D estimation will

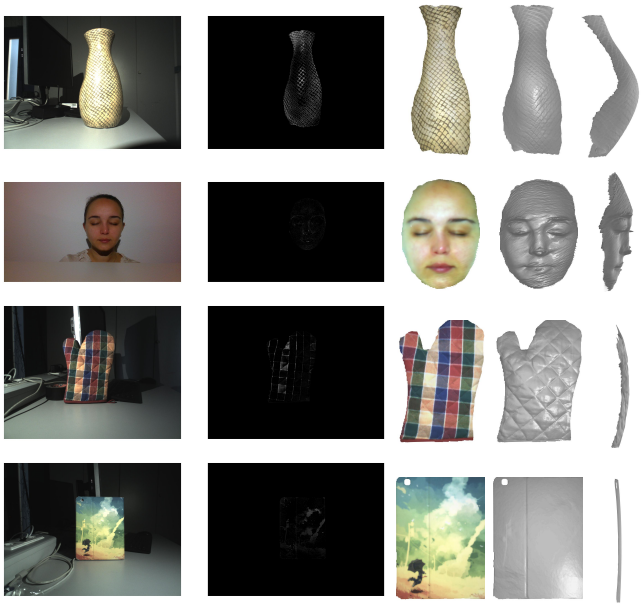

Fig. 3. **Qualitative evaluation on real datasets**. From top to bottom: *Vase*, *Face*, *Ovenmitt*, and *Tablet*. The same information is shown for the all cases. From left to right, it is displayed an arbitrary i-th input image, the corresponding i-th estimated specular map, the albedo estimation, our 3D estimation and an alternative point of view of that reconstruction.

be also more robust and accurate. Finally, and as a consequence of the above, the global method is more stable and faster in terms of computational cost, converging earlier and giving an speed up of $1.07\times$ and $1.06\times$ in comparison with UPS [11] and SAUPS [12], respectively.

4. CONCLUSION

In this paper an algorithm has been proposed to sort out the PS problem in an unsupervised, unified, uncalibrated, efficient and variational fashion under general lighting. As a result, the method can jointly estimate lighting and specularities of an object, its 3D reconstruction as well as its reflectance, all of them, from a set of color visual signals with no requirement of training data. To that end, our algorithm exploits a physical-aware formulation where a photometric constraint is combined with spherical harmonics lighting, perspective projection and a detail-aware minimal initialization that provides a fully interpretable solution. As a result, the method can handle a wide variety of object shapes and materials with unknown reflectances. Extensive experimental results on both synthetic and real datasets show the superiority of our joint estimation in comparison with state-of-the-art solutions, validating the effectiveness of our coding of details. Our future work is oriented to extend our formulation to scenarios with strong occlusions that produce more complex interactions between the object shape and the lighting.

5. REFERENCES

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

[2] 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.

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

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

[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] 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.

[7] Y. Queau, T. Wu, F. Lauze, J. D. 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.

[8] 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.

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

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

[11] 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.

[12] P. Estevez and A. Agudo, "Uncalibrated, unified and unsupervised specular-aware photometric stereo," in *ICPRW*, 2022.

[13] 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.

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

[19] B. Kaya, S. Kumar, C. Oliveira, V. Ferrari, and L. Van Gool, "Uncertainty-aware deep multi-view photometric stereo," in *CVPR*, 2022.

[20] S. Marschner and P. Shirley, *Fundamentals of Computer Graphics*, A K Peters/CRC Press, 2018.

[21] 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.

[22] A. Agudo, "Safari from visual signals: Recovering volumetric 3D shapes," in *ICASSP*, 2022.

[23] The Joyful Yell, ," URL: http://www.thingiverse.com/thing:897412.

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

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

[26] 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.