IEEE Open Journal of the
Computer Society

Received 12 April 2024; revised 6 June 2024; accepted 10 June 2024. Date of publication 27 June 2024;
date of current version 10 September 2024. The review of this article was arranged by Associate Editor Jingcai Guo.

Digital Object Identifier 10.1109/OJCS.2024.3419832

Non-Lambertian Surfaces and Their Challenges for Visual SLAM

SARA PYYKÖLÄ, NICLAS JOSWIG, AND LAURA RUOTSALAINEN

Faculty of Science, Department of Computer Science, University of Helsinki, PL 64 Helsinki, Finland

CORRESPONDING AUTHOR: SARA PYYKÖLÄ (e-mail: sara.pyykola@helsinki.fi).

This work was supported by the Research Council of Finland Flagship program: Finnish Center for Artificial Intelligence FCAI and the Research Council of Finland under Project 347197 and in part by the Artificial Intelligence for Urban Low-Emission Autonomous Traffic (AIforLEssAuto), Department of Computer Science, University of Helsinki.
The code is available via GitHub: https://github.com/helsinki-sda-group/SPINS
The SPINS dataset is available via Zenodo: https://zenodo.org/records/13140595

ABSTRACT Non-Lambertian surfaces are special surfaces that can cause specific type of reflectances called specularities, which pose a potential issue in industrial SLAM. This article reviews fundamental surface reflectance models, modern state-of-the-art computer vision algorithms and two public datasets, KITTI and DiLiGenT, related to non-Lambertian surfaces' research. A new dataset, SPINS, is presented for the purpose of studying non-Lambertian surfaces in navigation and an empirical performance evaluation with ResNeXt-101-WSL, ORB SLAM 3 and TartanVO is performed on the data. The article concludes with discussion about the results of empirical evaluation and the findings of the survey.

INDEX TERMS Monocular depth estimation, navigation, computer vision, non-Lambertian surfaces, specularities.

I. INTRODUCTION

Computer vision has since 1970s sought the visionary ideal of a machine that could see and perceive like a human being [1]. In spite of the optimism of the early days, the ultimate solutions have turned out to be elusive. One of the essential challenges of modern computer vision lies in *featurelessness*, which can be roughly translated as the lack of anything to be observed by image sensors, eg. in uniformly coloured walls or strongly reflecting surfaces called *non-Lambertian surfaces* [2]. The immediate consequence of featurelessness is the absence of useful quantifiable information, which may then disturb the functionality of computer vision algorithms as a whole.

The aforementioned non-Lambertian surfaces cause challenges especially to computer vision based navigation [3], [4], as the tasks of localization and image patch feature matching become increasingly difficult for the regions occupied with the reflections [2]. This challenge brings forth the risk of significant loss in accuracy and also the inevitable research question of how to effectively address the issue in terms of the algorithm and equipment. While challenges associated with non-Lambertian surfaces have long been recognized and considered in research, the rise of autonomous operations has thrust this issue back into the spotlight. Drones are now increasingly utilized in safety-critical operations, such as search and rescue missions, as well as for entertainment at large public events. As their usage grows, so does the prevalence of radio jamming attacks [5], necessitating a heavy reliance on computer vision for perception. However, computer vision systems are known to operate less reliably over water and highly reflective, transparent surfaces [6], which are examples of non-Lambertian surfaces. The ongoing expansion of terrestrial autonomous driving also grapples with challenges posed by non-Lambertian surfaces [7].

Our research ultimately aims at developing methods for industrial robots' autonomous indoor navigation using Simultaneous Localisation and Mapping (SLAM) and monocular cameras. As automation increasingly infiltrates industrial operations [8], discussions regarding non-Lambertian surfaces are also gaining relevance within this domain. Our goal is to develop a deep learning-based SLAM method with improved depth estimation, which is an essential step for accurate monocular vision-based navigation solution. The industrial environment includes various non-Lambertian surfaces however, which creates challenges for SLAM and depth estimation. Another complication emerges from the collection of the data: the industrial data is often confidential, and thus testing of different algorithms in industrial setting is practically impossible. We will address these challenges by a comprehensive survey into non-Lambertian surfaces' research involving multiple fields of computer vision, and concluding from the survey, propose a new outdoor dataset reflecting the expected properties of industrial environment. The proposed dataset will also give the possibility of exploring generality of different autonomous driving solutions, as the data is collected from a seaside promenade accessible to cars. Finally, we will perform an empirical study on the dataset and propose next steps for further research.

This article will start by presenting the fundamental surface reflectance models of non-Lambertian surfaces in Section II, and then continue on relevant studies from various fields of computer vision in Section III. Then in Section IV, we will review two datasets, KITTI [9] and DiLiGenT [10], used in the surveyed studies, and present a new dataset, SPINS, which is, as far as we know, the first one especially targeted for non-Lambertian surfaces' study in navigation. In Section V, we will perform an empirical evaluation with a modern state-of-the-art algorithm, ORB SLAM 3, on the new dataset to demonstrate its performance with non-Lambertian surfaces. The article is finished with discussion about the results, conclusions of the survey and recommended next steps for further research in Section VI.

The Sections II, III and IV are based on the corresponding author's master's thesis [11], which is a publicly available document but not a peer-reviewed publication. Sections V and VI regarding the new dataset and its empirical testing are novel and continued research.

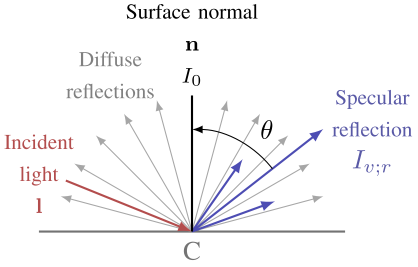
II. THEORETICAL BACKGROUND

The theoretical background for the proposed solutions so far regarding non-Lambertian surfaces, is a convoluted mix of knowledge spanning diverse topics and four decades of research [10], [12]. To better put the solutions into the context, the physical models related to the problem formulation are presented first briefly. For a complete treatise of the physics and physical quantities regarding the problem formulation, we refer the reader to McCluney's book [13].

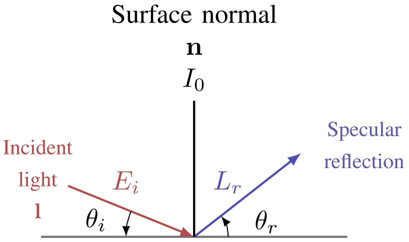
A. LAMBERTIAN REFLECTANCE MODEL

One of commonly used models to approximate light reflection is the Lambertian reflectance model, where a light hitting a matte surface is reflected *diffusely* according to Lambert's cosine law [14]. Mathematically formulated, the surface's reflected luminous intensity, denoted by $I_{v;r}$, is a function of θ, the angle between the two-dimensional surface normal $\mathbf{n}$ and specular reflection's direction; I_0, the reflected luminous intensity of the surface; and the surface colour C,

$$I_{v;r} = C \cdot I_0 \cdot \cos\theta. \tag{1}$$



(a) The Lambertian reflectance model.



(b) The BDRF model.

FIGURE 1. Two-dimensional diagrams depicting different surface reflectance models. The blue arrow marks the specular reflection, while the red arrow is the incident light. The gray arrows on the left mark the diffuse reflection. The vertical and thick black line is the surface normal and the other horizontal black line the surface level. *C* stands for colour, I_r for the reflection's intensity and *l* for the directional vector of incident light. E_i denotes the incident light's irradiance and L_r is the reflection's radiance. θ_i and θ_r mark the angles of incident and reflected light's direction with respect to the surface level.

In many contexts, the C as a constant is omitted however, as it only affects the scale of the reflected luminous intensity. An alternative formulation of the model is that the luminous intensity of the surface follows Lambert's cosine law and the luminance is isotropic, ie. uniform in all directions [14]. Then the model can be expressed in the vector form as

$$I_{v;r} = \underbrace{\langle \mathbf{n}, \mathbf{l} \rangle}_{\cos\theta} \cdot C \cdot I_0, \tag{2}$$

where $\langle \cdot\,,\,\cdot \rangle$ denotes dot product, $\mathbf{n}$ is the surface normal vector and $\mathbf{l}$ is the normalized incident light's direction vector.

This model has coined diffusely reflecting surfaces as Lambertian surfaces and their opposite as non-Lambertian surfaces with specular reflections, or shortly put, *specularities* [15]. Intuitively the visual difference between these can described as if the non-Lambertian specular reflection is concentrated on a random surface point, instead of being scattered evenly around the point where the light hits the surface as with the Lambertian diffuse reflection. The specular reflection therefore creates a lobe, a teardrop-shaped ray formation, bouncing off the surface, whereas the diffuse reflection creates a half-circle of rays, as illustrated in Fig. 1(a) [14].

While the model is attractive for its simplicity and ease of use, it is generally regarded as too inaccurate due to its failure in following conditions: i) the model can't account for bright surfaces' non-diffuse reflections and, ii) with appropriately large angles, surfaces are known to exhibit properties of both Lambertian and non-Lambertian surfaces [14]. As the model ignores these spectrums at large, a more intricate model is needed to quantify the difference in this situation. Thus, a bi-directional reflectance distribution function is often more accurate and is used instead, as the non-Lambertian reflectance is only a special case in its framework.

B. BI-DIRECTIONAL REFLECTANCE DISTRIBUTION FUNCTION

The bi-directional reflectance distribution function (BRDF) is a more general and fine-grained model, which can be used in two or three dimensions to model the reflected radiant energy in a given direction [15]. Mathematically put, it is the ratio of reflected ray's radiance, L_r, to the incident ray's irradiance, E_i, given their directions with respect to the surface normal $\mathbf{n}$, denoted as θ_r and θ_i and called the elevation angles, formulated as

$$f_r(\theta_r, \theta_i) = \frac{dL_r(\theta_r)}{dE_i(\theta_i)}. \tag{3}$$

The name for the model stems from the fact that the directions of incident and reflected ray can actually be reversed without changing the function's value [16], a property which is called the Helmholtz reciprocity [1].

In three dimensions, two more angle parameters called the azimuth angles, ϕ_r and ϕ_i, are added to measure the rays' direction with respect to the surface tangent (z-axis), forming a polar coordinate system of the form

$$f_r(\theta_r, \theta_i, \phi_r, \phi_i) = \frac{dL_r(\theta_r, \phi_r)}{dE_i(\theta_i, \phi_i)}. \tag{4}$$

The three-dimensional BRDF can additionally be parametrized with a half-angle bisector coordinate system, formulated as the following equation,

$$f_r(\theta_d, \theta_h, \phi_d, \phi_h) = \frac{dL_r(\theta_h, \phi_h)}{dE_i(\theta_d, \phi_d)}, \tag{5}$$

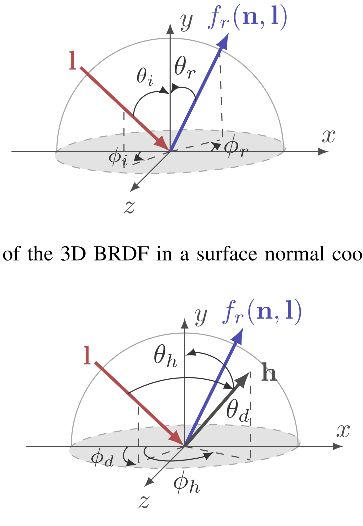
where $\mathbf{h}$ is the half-angle vector for the light rays, ϕ_h and θ_h are this vector's polar coordinates in the surface normal coordinate system and ϕ_d and θ_d are the incident light's polar coordinates in the transformed coordinate system. A 2D BRDF is illustrated in Fig. 1(b) and the 3D BRDF in each coordinate system is illustrated in Fig. 2. To simplify the setting, the coordinate system is normalized so that the x-axis is the surface tangent, then y-axis is the surface normal and the z-axis is the viewing direction, as is common in the computer vision context.

As the BRDFs are thus characterized by the incident light, surface normal and half-angle vector between the lights and their angles with respect to each other, it is common practice to present them in the following vectorized form, where $\mathbf{n}$ is the 3D surface normal vector, $\mathbf{v}$ is the 3D viewing point, $\mathbf{h}$ is the 3D half-angle vector and $\mathbf{l}$ is the 3D incident light vector [10]. Then the BRDF becomes

$$f_r(\theta_r, \theta_i, \phi_r, \phi_i) = f_r(\mathbf{n}, \mathbf{l}), \tag{6}$$

$$f_r(\theta_d, \theta_h, \phi_d, \phi_h) = f_r(\mathbf{v}, \mathbf{l}),$$

$$\mathbf{h} = \frac{\mathbf{l} + \mathbf{v}}{||\mathbf{l} + \mathbf{v}||},$$



(a) A diagram of the 3D BRDF in a surface normal coordinate system.

(b) A diagram of the 3D BRDF in a half-angle bisector coordinate system.

FIGURE 2. Diagrams of three-dimensional BRDF models in two different coordinate systems. The parameters of the BRDF are different depending on the coordinate system. Red indicates incident light, blue reflected light and black the half-angle vector. The gray circle represents the object's surface and the coordinate axes the surface level (*x*), normal (*y*), and tangent (*z*), respectively.

$$\theta_h = \langle \mathbf{n}, \mathbf{h} \rangle = \arccos\left(\mathbf{n}^T \mathbf{h}\right),$$

$$\theta_d = \langle \mathbf{l}, \mathbf{h} \rangle = \arccos\left(\mathbf{l}^T \mathbf{h}\right). \tag{7}$$

From now on, we shall use (6) and (7) to refer to the BRDF formulation, for the sake of brevity.

Further generalizations of the BRDF model have been presented. One of these is the spatially varying BRDF (SVBRDF), where two parameters are added to measure incident ray's location on the surface, so as to model the spatially varying surface normals. Another is a bi-directional surface scattering reflectance distribution function (BDSSRDF) with eight parameters, to quantify surface's internal scattering [13]. However, the true power of BRDF lies in the fact that the bi-directional function f_r can be defined case-wise, allowing flexibility over suitable functions and different cases [15]. For example, the Lambertian reflectance is a special case, where f_r is a constant function [10].

The main observation from the model is, however, that non-Lambertian surfaces' specular reflections are dependent on a viewing angle, as the angle parameters change depending on the viewing direction. Thus when the incident light is kept constant, and the viewing direction is changed — or equivalently, the non-Lambertian surface is tilted and the viewing direction and incident light are kept constant — a visible difference in brightness and colour of the surface can be observed. Consequently, the non-Lambertian reflectance violates the following assumptions: brightness constancy, where it is assumed that the brightness of pixels stays constant across images; and colour constancy, where it is assumed that the colour of objects stays constant across images [1].

Due to these violations, non-Lambertian surfaces can, for example, lead to false image patch feature matches or surface reconstruction results relying on texture, colour or brightness values, as specularities distort the values in one frame but not the other, depending on whether the viewing point and illumination stays fixed [17]. Furthermore, as various depth estimation methods and LiDAR sensors rely on sufficient texturing and Lambertian reflection for reliable depth info, the information can be erroneously estimated for non-Lambertian surfaces [4]. Another notable subclass of problems arises in medical computer vision, where local, sparse and strong specularities occur in closed non-adjustable environments of human bodies during medical endoscopies [10]. The mucous fluids in human body act as non-Lambertian surfaces obscuring the investigated tissue underneath, as seen in the studies by Meslouhi et al. [18], Mirko et al. [19], and Saint-Pierre et al. [20]. These far-reaching complications along with the computer vision's natural interest in optical exceptions has led to versatile research about non-Lambertian surfaces, encompassing numerous fields of computer vision.

III. COMPUTER VISION ALGORITHMS FOR NON-LAMBERTIAN SURFACES

The major field actively studying non-Lambertian surfaces since 1980s is *photometric stereo*, which tries to enable accurate 3D reconstruction of objects under different lighting conditions and the extraction of surface normal maps from objects [10]. As such, non-Lambertian surfaces pose an immediate problem, as the specularities block the information that is meant to be extracted. Available methods vary depending on the test setting of the study, which we shall present next.

A. PHOTOMETRIC STEREO

In photometric stereo, there are multiple adjustable light sources present in the environment [1], and the research question is generally formulated as modelling the specularities' behaviour under different lighting conditions [10]. Likewise the name "photometric stereo" [21], alludes to the multiple light sources present in the environment taking up the role of image sensors in a traditional stereo system. The usage of additional different sensors, such as intensity and image sensors, is not out of the question in the studies of photometric stereo, contradicting the term "stereo camera", which generally refers to the camera's characteristics only [22].

1) PROBLEM FORMULATION

To study photometric stereo, a mathematical-physical framework for reflectance must be assumed as the starting point. The most flexible theoretical model for modelling reflections is the BRDF, which was presented previously in Section II. Assuming this model in the form presented in (6), the mathematical formulation for the problem with *calibrated photometric stereo* is the following: given $\mathbf{I}$, the $m \times k$-matrix of m observed points in k lighting conditions, $\mathbf{L}$, the $3 \times k$-matrix of k observed three-dimensional lighting vectors $\mathbf{l}$, and a fixed viewing direction $\mathbf{v}^T = (0, 0, 1)$, we are trying to solve $\mathbf{N}^T$, the $m \times 3$ matrix of the three-dimensional surface normal vectors $\mathbf{n}$ in m points, from the following equation, where $\circ$ denotes the element-wise multiplication,

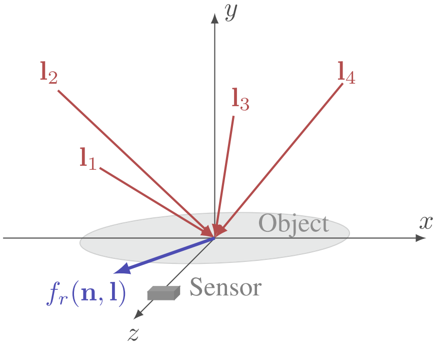


FIGURE 3. An illustrative diagram of a general test set-up in photometric stereo. Red arrows indicate incident lights and the blue arrow is the reflected light encoded by the BRDF. A dark gray box indicates the image or light sensor, and the light gray circle the object surface.

$$\mathbf{I} = \max\{f_r(\mathbf{n}, \mathbf{l}) \circ \left(\mathbf{N}^T \mathbf{L}\right), 0\}, \tag{8}$$

by using different assumptions and constraints on the $m \times k$-dimensional BRDF, $f_r(\mathbf{n}, \mathbf{l})$ [10]. Respectively in *uncalibrated photometric stereo*, the matrix $\mathbf{L}$ is unknown, and it needs to be estimated before solving the $\mathbf{N}$. The second argument of the max function, zero, represents the apparent shadow of the surface, which doesn't give up any information about the BRDF. The environment and test set-up is illustrated in Fig. 3.

2) LAMBERTIAN REFLECTANCE

The logical next question is what kind of different assumptions and constraints might be of use in solving the problem. One is naturally the Lambertian reflectance model, mathematically put as

$$f_r(\mathbf{n}, \mathbf{l}) \approx \mathbf{D}, \tag{9}$$

where $\mathbf{D}$ is a diagonal matrix with each row being a constant, representing the constant diffuse radiance [10]. If the $\mathbf{L}$ is known and has three different light vectors, the $\mathbf{N}$ can be uniquely solved by the linear least squares, and the reflectance values are the normalized rows of the $\mathbf{N}$ [21]. As this solution assumes the surface to be Lambertian and the linear least square is not able to reject the non-linear non-Lambertian surfaces, the accuracy suffers greatly from specularities [23]. Additionally it does not apply outdoors due to sun's nigh planar trajectory causing the inverse of the $\mathbf{N}$ to disappear. Finally, it is not truly useful in other realistic use cases either, as the assumption of Lambertian surfaces is rarely met sufficiently due to surfaces exhibiting both Lambertian and non-Lambertian properties [14]. Thus more recent efforts focus on a general unknown f_r that doesn't directly comply with the Lambertian reflectance model.

3) OUTLIER REJECTION METHODS

Another useful assumption for solving the problem would then be the local and sparse nature of specularities and shadows. In other words, they are spatially apart and occupy only a small portion of the image, even as groups. This gives us a possibility to detect and discard shadows and specularities as outliers. There exists various *outlier rejection methods* detection and rejection criteria [24]. For example, in studies by Verbiest et al. [25] and Wu et al. [23], outliers are modeled as Markov random fields due to their non-isolated grouped nature, thus capitalizing on the expectation maximization algorithms capable of optimizing the surface normals and realistic visual reconstructions of the surfaces. In contrast, a more recent avenue assumes the outliers to form a sparse matrix **E**, which is added to the Lambertian reflection matrix **D**, formulated as

$$f_r(\mathbf{n}, \mathbf{l}) \approx \mathbf{D} + \mathbf{E}. \tag{10}$$

Consequently by minimizing the rank of **E** with more elaborate statistical criteria — which translates as the reduction of linearly dependent noise present in the system of equations formed by f_r and **D** — we can achieve more robust rejection of specularities. An example of this approach can be made from [26], where a hierarchical Bayesian approximation is used to estimate surface normals while modeling the **E** and enforcing its rank to three based on the same rank of the surface normals and lighting vectors formulating the $m \times k$ -dimensional image, thus limiting the number of possible Lambertian reflections available for the image.

The inherent weakness of any outlier rejection method lies in the implicit assumption of local and sparse outliers: when met with large (global) specularities occupying a large part of the image (dense), the algorithms' accuracy decreases [10]. Additional problems arise from non-Lambertian diffusive surfaces that don't fit in the statistical model [26]. Other aspects worth considering are the computational complexity of the EM models regarding the fine-tuning of the parameters and amount of input images needed for reliable and valid statistical analysis. For example, a dense set of images means having over 100 images per reconstructed object [23], [25]. Hence, despite being robust approaches, the research in photometric stereo has veered towards *analytical BRDFs*.

4) ANALYTICAL BRDFS

Analytical BRDFS aim to solve the problem by accounting for outliers [10]. Some analytical BRDFs include methods modelling the surfaces as a set of microfacets, microscopic surface areas acting as individual specular reflectors. The distribution of microfacets' normals then differ from the surface normal depending on the surface's characteristics specified by the model. For example, microfacets can be assumed to be perfectly specular and thus only the microfacets with their normal equal to the half-angle vector **h** can cause specularities in the viewing direction [27]. This model has been adapted, for instance, in a study of uncalibrated photometric stereo [28]. In general though, more studies have been dedicated to the Ward model, which assumes an elliptical Gaussian distribution for the isotropic microfacet normals, thus having no preference over the reflectance direction [16]. The studies by Chung and Jia in 2008 [29], Goldman et al. in 2010 [30] and Ackermann et al. in 2012 [31] all use the Ward model or a derivation of it. The exact approach to the outliers in the Ward model varies. Some examples include shadows to estimate the parameters of the BRDF [29], the optimization of the object shape and model parameters, and pixel-wise parameters and surface normals in alternating turns [30], and finally, selecting the less shadowed pixels, which are most likely to offer viable info about the BRDF [31].

In the end though, while analytical approaches have the strength of accuracy on their side, little can be done about their weaknesses: the analytical models are material-specific [10], the models can be non-linear, requiring careful and long optimization and finally, there is no guarantee that an analytical model fits the observed BRDF well [15]. For example, while Goldman et al. observe that there are "fundamental materials", which make up most of the objects in real-life use cases and even constrain the materials' amount to two per object in their study, their assumption of linear combinations for the materials' BRDFs still leads to the task of solving a non-linear equation and estimating the surface normal at the same time [30].

5) GENERALIZED BRDFS

Another avenue in photometric stereo aims to overcome the challenge of generalizable BRDFs by using the general properties of BRDF, such as monotonicity, Helmholtz reciprocity and isotropy [10]. These *generalized BRDFs* are further supported by the fact that materials often show structured BRDF values in real life, implying isotropy. Isotropy simplifies the mathematical formulation of the BRDF in a half-angle coordinate system, presented in (7): the function has now only three parameters, as the ϕ_h is no longer necessary [32]. Monotonicity, in the other hand, implies that the intensity increases as the input increases in value, giving a unique inverse function for the BRDF. These constraints open various possibilities, including *the bi-polynomial approximations* as various models don't anymore show significant dependency on ϕ_d either, as demonstrated in [15]. The bi-polynomial model is then formulated in the following fashion as

$$f_r\left(\mathbf{v}, \mathbf{l}\right) \approx g\left(\theta_h, \theta_d\right). \tag{11}$$

A bi-polynomial model with the assumption that the aforementioned equation can be factored into two separate terms $g_1(\theta_h)$ and $g_2(\theta_d)$ is used by Shi et al. in their study [33]. This model enables the iterative estimation of the surface normal in a suitable slow-varying low-frequency domain with shadow and specular cut-off thresholds. Another generalized BRDF without the bi-polynomial model assumes the incident light to be collocated with the viewing point, allowing to decouple the surface normal from the BRDF [32]. These generalized approaches bring reasonable approximations of multitude of reflections with various computational complexities, but they have difficulties dealing with anisotropic reflections, which remains as an actively studied challenge [10].

Another avenue implicitly used both in generalized and analytical BRDFs is *the component-wise structure of the reflection* [15]. In this case, the overall reflection is built as a sum of two or three separate components, such as mirror, specular or diffuse reflection components, which can individually accommodate different reflectance models [14]. An example of this approach is the use of a pseudo-specular and diffuse Lambertian component. These component models tend to be approximations of the actual mappings, and thus some loss of accuracy is often present, where the observed reflectance doesn't fit the components directly.

6) NEURAL NETWORKS

The most recent venture in photometric stereo involves using *neural networks* that learn the reflectance mapping directly [39]. While this approach tends to have the advantage of dealing with majority of reflectance at ease, a few limits exist: significant error with noisy light intensities, long runtime and decreased performance in complex reflections [41] and problems with concave shapes [40]. Another notable neural networks study is nLVMS-Net [52], which uses a shape-from-shading network to formulate probability densities for per-pixel surface normals in each view and finally volume cost filtering to learn the true per-pixel surface normal and also depth.

Another avenue involving neural networks has a different approach. This approach investigates how neural networks or machine learning could be used in tandem with physical models, bringing together the best of the two worlds: the accuracy and explainability of physical models and flexible and robust computations of neural networks. The first results are already available, for example by Geourgoulis et al. with a CNN learning the reflectance map of single non-Lambertian material and then fitting parameters to it [42], Li et al. with a SVBRDF-modeling CNN [43], and Rhee and Lee with a GAN-guided CNN [44]. For a multi-spectral setting, thus accounting for wavelength as well, there is a study by Lv et al. [53]. Similar ideas are used for the purpose of object reconstruction by Chen et al. [54] and Sang et al. [45], while Iwaguchi and Kawasaki [55] optimize the training phase of photometric stereo using DNNs.

7) UNCALIBRATED PHOTOMETRIC STEREO

As noted earlier, uncalibrated photometric stereo has to estimate the lighting matrix $\mathbf{L}$ in along with the surface normal matrix $\mathbf{N}$ Equation (8) [10]. Mathematically, uncalibrated photometric stereo is based on the assumption that the reflectance is Lambertian so that the albedo-scaled lighting matrix $\mathbf{L}$ and surface normal matrix $\mathbf{S}$ formulate the observed image [56]. As the normals are scaled, we can solve the ambiguity caused by scaling, denoted as $\mathbf{A}$, by the singular value decomposition or matrix factorization [57]. Thus the whole problem can be stated as the following equation,

$$\begin{aligned}\mathbf{I} &= \max\{\mathbf{D} \circ \left(\mathbf{N}^T\mathbf{L}\right), 0\} = \mathbf{S}^T\mathbf{L} \\ &= \hat{\mathbf{S}}^T\mathbf{A}^T\mathbf{A}^{-1}\hat{\mathbf{L}}. \end{aligned} \tag{12}$$

where $\hat{\mathbf{S}}$ and $\hat{\mathbf{L}}$ are the unscaled pseudo-normal and pseudo-lighting matrices respectively [56]. However, solving the final ambiguity matrix $\mathbf{A}^T\mathbf{A}^{-1}$ requires additional steps in the testing or calibration phase, which is an intricate and onerous process [10]. For instance, the rotation ambiguity can be solved with six different surface points with constant albedo or intensity, but more common constraints are the integrability of the surface or the observation of shadow boundary. These reduce the problem to *the Generalized Bas-Relief ambiguity*, stated as

$$\begin{aligned}\mathbf{I} &= \max\{\mathbf{D} \circ \left(\mathbf{N}^T\mathbf{L}\right), 0\} = \mathbf{S}^T\mathbf{L} = \hat{\mathbf{S}}^T\mathbf{A}^T\mathbf{A}^{-1}\hat{\mathbf{L}} \\ &= \hat{\mathbf{S}}^T\mathbf{G}^T\mathbf{G}^{-1}\hat{\mathbf{L}}. \end{aligned} \tag{13}$$

For solving the matrix $\mathbf{G}$ with three unknown variables, there are numerous alternative solutions to choose from, such as the perspective camera model, a ring of light sources or an analysis of the specularities. Each approach comes with their own limitations and advantages. For contrast, chromatic clustering can be used to detect points, which have equal albedo [46], locating the points where $\mathbf{n} = \mathbf{l}$ [47], and minimizing the entropy after assuming a limited amount of dominant colours in the image [48]. The first method is unsuitable to grayscale images, the second is limited solely to the diffuse component and the third requires intricate pre-processing steps. Recently, neural networks have also been utilized to solve the uncalibrated problem in a general form without the assumption of the Lambertian reflectance or the uniform distribution of light sources required for solving general BRDFs, in a computationally costlier manner [49].

A notable study using neural networks in uncalibrated photometric stereo is the LCNet [49], which uses convolutional layers and max pooling to detect global features from local features. The network has been further enhanced in the studies by Chen et al. [39] and Kaya et al. [40], ranking it as a state-of-the-art system in photometric stereo. However, it still has performance issues with ambiguous special cases, such as piece-wise planar surfaces, planar surfaces with uniform albedo and concave shapes.

Another avenues outside the Equation (12) are *the manifold embedding methods*, which acquire the surface normals up to a rotational ambiguity and then use additional constraints, such as integrability or shadow boundary, to solve that [10]. A few examples include the studies by Sato et al. [50] and Lu et al. [51].

8) OTHER APPROACHES

Finally, aside from solving the (8) and (12), there are numerous solution methods involving the change of the data collection method or the input [10]. For example, a multi-spectral light field may be used, which gives additional constraints of multiple viewpoints and point lights and thus more accurate measurements of the surface normals' orientation [36]. Object motion works in a similar albeit stricter manner [34], [35]. An alternative solution can be found from the colour channels, which reveal specularities when studied

TABLE 1. A Table Describing Briefly the Proposed Solutions in Photometric Stereo

Problem formulation	Proposed solution	Studies
Calibrated photometric stereo	Lambertian reflection and linear least square	[21]
	Statistical non-Lambertian outlier rejection	[23], [25]
	Analytical BRDFs	[28], [29], [30]
	Generalized BRDFs	[14], [32]
	Bi-polynomial approximations	[33]
	Specialized data collection methods	[34], [35], [36]
	Depth priors	[37], [38]
	Masking methods	[18], [19], [20]
	Neural networks	[39], [40], [41]
	Hybrid models	[42], [43], [44], [45]
Uncalibrated photometric stereo	Chromatic clustering and equal albedo	[46]
	Collocated $\mathbf{l}$ and $\mathbf{n}$	[47]
	Minimizing entropy	[48]
	Neural networks	[49]
	Manifold embedding methods	[50], [51]

individually. This approach has also been utilized to create sophisticated *masking methods* for medical endoscopies, [18], [19], [20]. Another interesting method is the use of *depth priors:* by fusing a priori depth info as regularizers for the final reconstructions, corrections can be made in problematic low-frequency domains, [37] [38]. Lastly, other alternative solutions to photometric stereo include also colored lighting, a perspective camera model, which is more accurate than the traditional model, and cameras with non-linear response, among many other methods. While these solutions have generally gained interesting results, their weaknesses tend to be often the lack of robustness or the specialized data collection method, which is often not practically feasible, easily adjustable and possibly not even usable outdoors or outside any controlled environment.

This concludes our survey on photometric stereo. All presented studies are summarized in Table 1. As we have now gone over photometric stereo in an extensive manner, we turn our sights into navigation, where non-Lambertian surfaces play a significant role as well.

B. NAVIGATION

Computer vision research of autonomous robot navigation started at 1980s [1]. The research has progressed rapidly in various areas, producing a new golden standard for today: *Simultaneous Localization And Mapping* (SLAM) [12]. SLAM solves localization ("where we are") and mapping ("what is around us") at the same time. Consequently, the availability of different sensors has been a major driving force in the emergence of new SLAM methods. Regardless of the used sensors however, non-Lambertian surfaces pose a veritable challenge for SLAM via its problem formulation, as is demonstrated next.

1) PROBLEM FORMULATION

The most basic mathematical formulation of the SLAM is probabilistic and consists of state, observation, control input and landmark vectors [58]. The state vector $\mathbf{x}$ describes the pose, location and orientation of the sensor at a state t. Spanning t discrete states, the control input vector $\mathbf{c}$ defines the transition from each state to the next in the system, for example the speed and direction of the sensor system. An i-dimensional vector of observed landmark points $\mathbf{l}$ along the way is recorded from the t-dimensional observation vector $\mathbf{o}$, which includes for example camera images or inertial sensor measurements. We then need to compute the following joint probability distribution using the Bayes theorem,

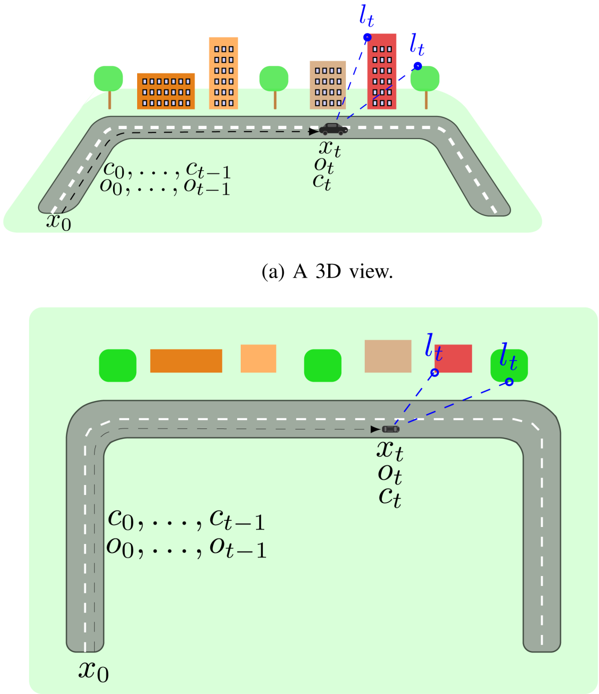
$$P(x_t, \mathbf{l}|\mathbf{c}, \mathbf{o}, x_0). \tag{14}$$

This task is solved via updating the prior distribution of current state, and the posterior distribution of measurements, deriving accurate observation and motion models for the sensor system in the process. The determination of motion model is referred to as *localization*, which can performed in various ways. When it is done by computing the motion between consecutive images, it is referred to as *visual odometry*. In contrast, *mapping* is the process of creating a map of the environment to help with the localization, presented as the vector of landmark points. The problem formulation is demonstrated in Fig. 4.

2) APPROACHES

General approaches how to construct these observation and motion models can be categorized based on the sensors and input data [12]. An alternative categorization of SLAM is between *indirect* (feature-based) and *direct methods* based on their general method of solving the problem. The indirect methods use only features derived from the observations and their respective descriptors, whereas the direct methods use all of the observed data. Another categorization can be done based on the density of the environment maps, as *dense* or *sparse environment maps* may be sought [58]. We shall use the split of direct and indirect SLAM methods from now on, as it provides the most relevant framework of inspecting non-Lambertian surfaces' and their challenges.

Indirect SLAM methods detect features, match those between subsequent images and track them over image sequences, which is called *tracking* [12]. Let us now observe that non-Lambertian surfaces are ideal to cause problems for feature-based tracking, as specularities can lead to a loss of tracking information via false or missing feature matches. This is due to colour and brightness constraints violated by the non-Lambertian surfaces, as these characteristics can play an important role in feature detection.



(a) A 3D view.

(b) A 2D view.

FIGURE 4. A two diagrams illustrating the SLAM problem in an urban setting, prevalent in applications for autonomous navigation. The black dashed line with an arrow marks the trajectory of the black car, and the blue circles and dashed lines mark the observed landmark points l_t. The control inputs (here driving cues) at a time step t are denoted by c_t, whereas the observations (eg., images) are denoted by o_t and the car's current location by x_t and the initial location by x_0.

However, non-Lambertian surfaces are not simple trivialities for direct methods either. Both direct and indirect SLAM systems can correct their mapping and localization when entering a previously visited area in a process called *loop closure*. Loop closure takes place to prevent *drifting*, a gradually occurring deviation between the actual and predicted location of the sensor system [59], resulting in better performance in the localization task. In the case of enough drifting, a loop-closure might be skipped and the drifting continues to grow, decreasing the localization accuracy. Hence non-Lambertian surfaces are a notable and diverse issue in navigation as well.

Finally, one last research question related to navigation and non-Lambertian surfaces at large is the research regarding LiDAR depth measurement accuracy of the non-Lambertian surfaces, such as water. We shall not address this question here and instead we shall refer the reader to Paul et al. for that [60].

Having summarized the main problem formulation, we can move on to relevant studies. To highlight the evolution of SLAM systems, we shall primarily proceed chronologically through the survey.

3) SLAM SYSTEMS

Often heralded as the first semantic SLAM [36], SLAM++ was published in 2011 [61]. While the preceding SLAM systems had been mapping sparse point features and other geometric primitives from the environment, SLAM++ opted for an object-oriented approach constructed on top of geometric primitives. Using KinectFusion and LiDAR sensors [62] to build an object database, pixels in the picture could be classified into different objects. This process is generally referred to as semantic segmentation nowadays [36], thus terming SLAM++ as semantic SLAM. In contrast, the concept of obtaining prior information of objects in the form of parameter vectors is nowadays referred to as shape priors and is still widely used in fields of SLAM and object detection [63], [64], [65], [66]. Semantic SLAM systems mark also a very prominent avenue for producing methods capable of coping with non-Lambertian surfaces, as the shape priors can fill in the info that specularities block from the images. Thus a few more notable studies are presented later on, which can be regarded as the current state-of-the-art or the enabling basis for them.

ORB-SLAM2 from 2017 [67] is even today considered as the state-of-the-art of indirect SLAM system due to its enhanced optimization and loop closure techniques, the wide range of applicable input from monocular and stereo images to RBG-D data [12], and a modern feature matching algorithm "Oriented FAST and rotated BRIEF" (ORB) which is built on the preceding algorithms of FAST and BRIEF, but with accelerated speed and better rotation invariance [68]. In ORB-SLAM2 the tracking, local mapping and loop closing are separated into their own threads, and the global bundle adjustment and motion optimization is performed only after the threads are completed. The drawbacks of this widely used lightweight open-source solution stem from the weak robustness for motion blur and featureless regions, which accumulate drift considerably in monocular input [2].

The next relevant milestone was the emergence of neural networks in SLAM systems around 2017 [12]. One of the first precursors was the CNN-SLAM, which utilized convolutional neural networks (CNN) with the ResNet architecture to perform semantic segmentation, and predict depth densely even in featureless regions by assuming a baseline stereo and then refining the keyframe depth maps with the baseline stereo, regularizer and each new frame's depth estimations and depth uncertainty maps [69]. CNNs continue to be widely used in semantic segmentation today [70]. DeepSDF [71] uses feed-forward networks in a probabilistic auto-decoder architecture to learn continuous signed distance functions (SDFs), which output the distance of a point coordinate x to a surface of interest.

With the decision boundary of f_d the surface of interest can be constructed via ray-tracing or the marching cubes algorithm [71]. By the universal application theorem, the feed-forward networks in DeepSDF are harnessed to approximate this function up to a computationally feasible precision with the loss function of $\mathcal{L}(f_\theta(x), s) = |f_c(f_d(x), \delta) - f_c(s, \delta)|$, where f_θ is the approximation of f_d produced by the network, defined by its parameter vector θ; δ is the control distance parameter to the surface of interest maintaining the metric SDF; and f_c is the real-valued "clamp function", defined as

$f_c(x, \delta) := \min(\delta, \max(-\delta, x))$. Using this loss function to learn the latent low-dimensional variables of surfaces, which can be directly inputted into the auto-decoder to be further optimized via back-propagation, gives a possibility to model varying SDFs. While admittedly DeepSDF is not a SLAM system, but a 3D object detection and reconstruction system, it is a vital part of a notable semantic SLAM system later on, which is why it is presented alongside other SLAM systems.

CubeSLAM, also a semantic SLAM, unites the fields of monocular 3D object detection and SLAM systems [64]. By feeding the pose estimation info of SLAM system to the object detection and the object detection information in turn to the SLAM pose and scale estimation, the benefit is mutual and amplifies both systems' performance. Aside from this symbiotic info recycling, the novelties of CubeSLAM lie in the mathematical approaches to the bounding boxes, measurement functions between objects, cameras and points and lastly memory efficiency to storing the objects. Two other semantic SLAMs addressing the non-Lambertian challenges are NodeSLAM [72] and From Detections to 3D Objects (FroDO) [73]. FroDO uses the DeepSDF and encoder architecture to further refine the monocular object detection via shape priors and other estimation steps, whereas NodeSLAM uses RGB-D data to optimize the embeddings with the help of a new rendering volumetric function, which needs fewer measurements and is capable of dealing with occlusion.

D3VO, a monocular visual odometry system [74], brings together the lessons of photometric stereo and SLAM united under brightness affine transformation and deep learning. Assuming

1) an affine brightness transformation due to a change of camera exposure, with $I_{t'}$ being the new changed intensity and I_t is the previous unchanged intensity, defined as $I_{t'} = aI_t + b, a > 0, b > 0$
2) a photometric error with comparison functions l, c and s for luminance, contrast and structure [75], defined as

$$r(I_t, I_{t'}) = \frac{\lambda}{2}(1 - \text{SSIM}(I_t, I_{t'})) + (1 - \lambda)\,||I_t - I_{t'}||_1\,,$$

$$\text{SSIM}(I_t, I_{t'}) = [l(I_t, I_{t'})]^\alpha \cdot [c(I_t, I_{t'})]^\beta \cdot [s(I_t, I_{t'})]^\gamma\,,$$

$$\alpha > 0, \beta > 0, \gamma > 0, 0 < \lambda < 1,$$

and thirdly

3) an uncertainty map of true pixel intensity y with Laplacian noise, defined as

$$-\log p(y|\hat{y}, \sigma) = \frac{|y - \hat{y}|}{\sigma} + \log\sigma + C, C \in \mathbb{R},$$

and embedding all three of these equations into a self-supervised re-projection error, D3VO can then adjust the weighting of the residual for areas with high uncertainty, which in turn achieves greater robustness against featurelessness.

The final state-of-the-art semantic SLAM we survey is DSP-SLAM [76]. DSP-SLAM uses ORB-SLAM2 architecture for sparse tracking and mapping and DeepSDF for the shape embedding, to produce sparse backgrounds and dense shape reconstructions via deep shape priors as its environment mappings. The input data can be monocular or stereo, the latter optionally with LiDAR, while the system runs at 10 frames per second. Being a sequential SLAM with both local feature and global object optimization, it differs from FroDO's batch implementation and NodeSLAM's local optimization based on depth images, but borrows inspiration from both. In the end, it brings about considerably good visual results in a low frame rate of 10 Hz.

We are going to close the survey by regarding the modern-state-of-art visual SLAM and odometry systems, who have touched on the problem of non-Lambertian surfaces especially. Firstly there is the direct and sparse odometry system, DSO, by Engel et al. from 2017 [77], who opted for photometric error around each pixel instead of using a geometric smoothness prior for all of the image data. Using this approach, Engel et al. report increased robustness and accuracy compared to other state-of-the-art systems, such as ORB SLAM 2. They mention that using the photometric error instead of geometric error in the is the cause for increased robustness and that no significant issues with non-Lambertian surfaces were present. However, most of the specularities are present in high-gradient points, where other information is aplenty. In the end, DSO fails under significant lighting changes, such as the time of the day or the direction of the light.

DynaSLAM, a visual SLAM system by Bercos et al. from 2018, builds on ORB SLAM 2 by adding dynamic object detection and background inpainting [78]. Resulting from motion segmentation, the tracking accuracy is decreased in KITTI's less textured sequences, where the reference points are far away and static. A continuation of this study as Dynamic-SLAM by Xiao et al. [79] in 2019, a semantic monocular SLAM. Combining an single-shot object detector based on CNN and prior knowledge to a feature-based SLAM pipeline, the system achieves increased tracking accuracy. The deep learning remains yet to be incorporated into backend, however, and the object detection is accelerated with GPU to achieve real-time performance.

Another study from 2019 is the visual SLAM by Dong et al. [3]. By using a novel feature matching system and traffic signs as landmarks, the system is able to increase the localization accuracy even in low-textured environments. However, the system reports weaker accuracy and robustness on initialization due to the use of regression trees in the process.

From 2021, there is the visual and direct SLAM system, DSV, by Mo et al. and a continuation of DSO [80]. Using LiDAR sensors, they extend the DSO framework by optimizing the 3D points' scale and then utilize a point descriptor for more efficient loop closures. The system does not incorporate IMU signals and thus uses the assumption of forward-moving camera. In the evaluation, the system failed in direct sunlight and other sudden brightness changes.

TABLE 2. A Summary Table of Presented SLAM and Odometry Systems

Studies	Approach	Density
SLAM++, [61]	RGB-D, semantic	Sparse
CNN-SLAM, [69]	Monocular, visual direct	Dense
CubeSLAM, [64]	Monocular, visual feature-based, semantic	Sparse
NodeSLAM, [72]	RGB-D, semantic	Dense
FroDO, [73]	Monocular, visual feature-based, semantic	Dense
DSP-SLAM, [76]	Monocular or stereo, LiDAR optional with stereo, feature-based, semantic	Dense objects, sparse environment
ORB-SLAM2, [67]	Monocular, visual feature-based	Sparse
D3VO, [74]	Monocular, visual feature-based	Dense
DSO, [77]	Monocular, visual direct	Sparse
DynaSLAM, [78]	RBG-D, indirect	Dense
DynamicSLAM, [79]	RBG-D, indirect	Dense
DSV, [80]	Monocular, direct visual	Sparse
Dong et al., [3]	Monocular, feature-based visual	Sparse

This concludes our survey on navigation. All presented studies are summarized in Table 2.

C. OTHER FIELDS

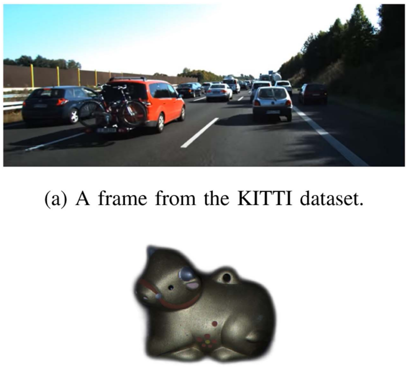
As we have now covered both photometric stereo and SLAM in the context of non-Lambertian surfaces, it is natural to move on to a field that is a synthesis of SLAM and photometric stereo, called *fusion methods* [81]. The idea of fusion methods is to use photometric stereo to define the surface normals, which can then be used to recover depth information needed for more fine-grained object reconstruction. Recent related studies include exploring different penalty functions for converting the surface normals to a depth estimate by Antensteiner et al. from 2018 and the usage of neural networks simultaneously to predict surface normals and depth by Zhan et al. [82] from 2019.

The approach of fusion methods nevertheless heralds an interesting approach to the question of depth estimation. Surface normals and gradients have been proven to be a viable source of depth information in studies by Zhang et al. [83] and Joshi et al. [84], and photometric stereo can be regarded as the most versatile and resourceful field regarding the estimation of surface normals. The major questions in utilizing photometric stereo techniques in depth estimation lie in the problematic special cases and error margin. A small angular error in surface normals may lead to a significant depth error [81], and secondly, there is the question of penalizing the edges and planar surfaces so that the depth estimation remains accurate in both cases. One difficult aspect is the computational complexity driven by this approach, as it has not been accurately commented in the papers. In other hand, while traditional and less costly BRDF models have been accurately fitted for a variety of materials and even general classes of materials [10], the neural networks so far have been most robust and general models with the expense of computation. Thus, a very promising avenue lies in the union of analytical BRDF models and neural networks, as they can account for each other's weaknesses.

Finally in our survey, we shall consider a multi-faceted field of computer vision called *illumination invariance*. The research of this topic covers diverse tasks, such as object detection [85], SLAM [86], and face recognition [87], but the essential question remains roughly the same across them: how can we extract the same information under illumination changes. While the question may very well seem to cover the specularities, the field is more focused on seasonal, daily or situational human-made changes in illumination rather than specularities, which are a direct effect of the prevailing illumination while not being a source of illumination or a change in the sources per se. As such, the methods are not particularly tested against non-Lambertian surfaces and further studies are required to determine whether the illumination invariant techniques would be fitting to cope with specularities.

IV. DATASETS

In the preceding section, we covered many existing methods regarding specularities. A crucial key question in this survey was how good exactly are these presented methods, but additionally, we must now ask what are the datasets used for these methods. If the used datasets don't show specularities properly, the methods' performance and applicability can rightfully questioned. Thus, it is time to cast a critical look



(a) A frame from the KITTI dataset.

(b) A picture of the object "Lamb" from the DiLiGenT dataset.

FIGURE 5. A frame from then KITTI dataset (a) and a picture of the object "Lamb" from the DiLiGenT dataset (b). The specularities in KITTI are soft, local and sparse, centered around cars' corners and windows. The specularities in DiLiGenT are dense but not rather strong, and local as well depending on the object.

into the most common datasets, KITTI in navigation [9] and DiLiGenT in photometric stereo [10]. Other datasets in these fields, such as TUM-monoVO [88] and Malaga [89], are not considered, as they are not standardised benchmarks for the field or have not been adopted into as general use due to various reasons, like poor availability, limited scope of input data or weak applicability to different navigation tasks. For example, both TUM-monoVO and Malaga utilize grayscale images. As such, while there could be specularities present in less used datasets, they do not offer viable information for the solution of SLAM nor research regarding the non-Lambertian surfaces, which are the center focus of this survey.

A. KITTI

KITTI was first recorded in countryside of Germany, in 2012, primarily for the research of autonomous driving [9]. The dataset can be loosely described as a continuous feed of images corresponding to views of a driver driving around a medium-sized city and its highways on a clear sunny day, as can be seen from Fig. 5(a). Specularities in a such setting are rather unequivocally sparse, local and soft, appearing only as few slightly lighter pixels focused on cars and other objects comprised of common non-Lambertian surfaces, such as glass and metal. Consequently due to KITTI's popularity [12], SLAM systems at large are not extensively researched regarding global and dense specularities. In contrast, only a rare few systems have actively paid attention to local and sparse featurelessness in their approaches to the best of authors' knowledge, even though the issue is of utmost importance in industrial SLAM.

B. DILIGENT

The second dataset to be inspected is DiLiGenT, a benchmark dataset in photometric stereo [10]. While DiLiGenT, unlike KITTI, was specifically designed with specularities in mind, the special settings presiding in the dataset are worth noting. DiLiGenT consists of dozens of static images taken from a limited set of objects with a same viewpoint and varying lighting, with the background completely blacked out, as illustrated in Fig. 5(b). The purpose is to assist the detection of specularities and to block out any noise regarding the measurement of light with finicky equipment, but at the same time it represents unnatural and highly ideal conditions for both photometric stereo and navigation. Furthermore, the specularities in DiLiGenT are mostly local, dense and soft in nature, being relatively pale and small areas in the surface of the object. Thus they do not properly account for global, strong and dense specularities, blinding and large reflections taking up a large portion of the image in total. Consequently we propose an even more challenging benchmark with such specularities in the next section, designed especially for use in navigation.

C. SPATIOTEMPORAL IMAGES FROM NON-LAMBERTIAN SEA (SPINS)

To demonstrate the impact of non-Lambertian surfaces and associated reflections on monocular depth estimation and visual Simultaneous Localization and Mapping (SLAM), we recorded a novel data sequence of 53,207 images by the name of "SPatiotemporal Images from Non-Lambertian Sea" (SPINS). This data was compiled during winter at a coastal seaside located in Helsinki. In addition to reflections emerging from the water, the presence of snow and ice creates further complications as they exhibit non-Lambertian properties. The data acquisition process was conducted on a winter morning shortly after sunrise to encapsulate the reflections occurring when the sun is relatively low, inducing an unfavorable reflection angle, particularly as the reflections directly interact with the camera.

To gather the image data, we utilized two widely available sensors: an iPhone 13 and a RealSense D455. Both devices were temporally synchronized via Network Time Protocol (NTP) during the entire experimental procedure. The RealSense images are recorded in resolution 1280 × 720 and 30hz framerate. The iPhone captures HD pictures, but we scale them down to 640 × 480 pixels before saving them to disk to avoid a disk overload. To acquire the reference trajectory for visual SLAM, the system was supplemented with a tactical grade SPAN system, renowned for its post-processing accuracy up to centimeter level. It contains a GNSS-Inertial fusion solution that is aided by Precise Point Processing (PPP) in postprocessing. We endeavored to capture the ground truth depth using LiDAR, but the challenge of determining depth towards water remains an unresolved issue within LiDAR research [60]. Consequently, we refrained from utilizing the LiDAR-measured ground truth for evaluation purposes. The whole system is carried by a pedestrian, so that the cameras are approximately on chest height pointing forward.

We aimed at quantifying the amount of non-Lambertian surfaces across the SPINS dataset by converting color images to gray-scale and thresholding at a value of 252 to detect bright spots as proxies for reflections. The results of this detection method are presented in Table 3. Due to the nature

TABLE 3. Specular Reflection Quantity Measures for the SPINS Dataset

Number of images	53 206
Number of images with specularities (%)	26 101 (48%)
Number of images where > 1% pixels are occupied by specularities (%)	7555 (14%)
Mean amount of pixels with specularities	0.4%
Mean amount of frames where at least 1 pixel is occupied by specularities	0.8%

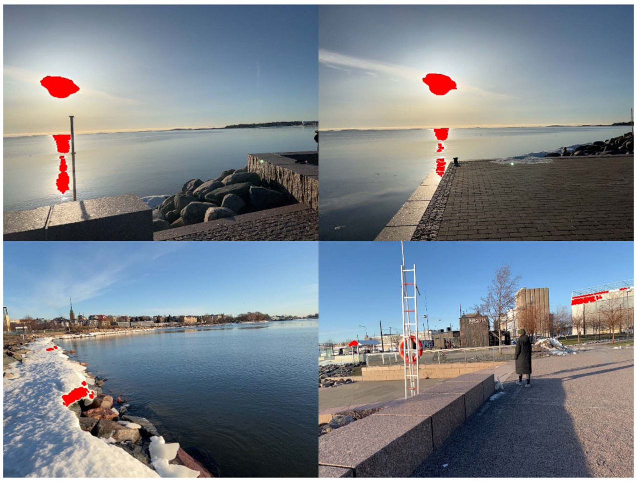


FIGURE 6. Illustrative samples showcasing reflection masks, with detected reflective areas highlighted in red for clarity.

of auto-exposure, extremely bright pixel values can be found in almost all images, but if no real reflection is visible, the detected spots are generally very sparse outlier pixels. To address this phenomena, we refined our approach by applying a morphological opening operation to the identified bright regions, followed by filtering to retain only those areas comprising of at least 20 pixels. This methodology effectively minimized the impact of outlier pixels, enhancing the detection of genuine reflective surfaces. The compiled statistics reveal that 48% of the images contain specular reflections, with 14% having more than 1% of their pixels reflecting. On average, specularly reflective pixels constitute 0.4% of an image, and 0.8% of images present at least one specularity, offering a nuanced view of the prevalence of non-Lambertian reflections within the dataset.

To illustrate the application of our reflection detection methodology, Fig. 6 displays four representative images with an overlaid reflection-mask. The scenes depict various environments, including waterfronts, a pedestrian pathway, and a snow covered path. The red overlay specifically highlights regions of significant brightness, which correspond to strong and dense specularities from the sun and water, as well as softer sparse specularities from various non-Lambertian surfaces, like windows and snow covered areas.

V. COMPARISON AND RESULTS

A. METRICS

We use Absolute Pose Error (APE) to evaluate the visual odometry VO system's accuracy by measuring the discrepancy between predicted poses and ground truth poses at each timestamp. To ensure a meaningful computation of the APE, the predicted trajectory needs to be aligned first with the ground truth trajectory using the Umeyama alignment method [90]. This method performs a least-squares optimization to minimize the differences between corresponding points in the two trajectories. By aligning the trajectories, the APE provides a reliable assessment of the true error in pose estimates, thus offering a comprehensive evaluation of the VO system's performance.

B. MONOCULAR DEPTH ESTIMATION

Estimating depth from monocular images poses a significant challenge, as it requires recovering the 3D structure of a scene from a single 2D image. This process is inherently ambiguous due to the loss of depth information when projecting the 3D world onto a 2D plane. Consequently, distinguishing between objects that are small and near the camera and those that are large and far away, as well as inferring relative distances between objects in the scene, becomes difficult. This problem is further compounded by factors such as varying lighting conditions, reflections, and textureless surfaces, which can introduce additional complexity to depth estimation.

Monocular depth estimation methods can be broadly classified into two categories: supervised and self-supervised. Supervised methods rely on labeled ground truth depth data for training, whereas self-supervised methods exploit the inherent structure in the data, such as image sequences or stereo pairs, without the need for explicit depth annotations. Supervised methods learn the absolute scale from ground truth depth data but often depend on strong assumptions, such as the camera being at the same height during inference as it was during training. In contrast, self-supervised methods require additional information or assumptions to recover the true scale. However, they offer greater flexibility during both training, where differently annotated data can be fused, and inference, as they predict relative depth.

To demonstrate the impact of non-Lambertian surfaces on monocular depth estimation, we employed the state-of-the-art model from Ranftl et al. [92], ResNeXt-101-WSL, which is trained on a large variety of datasets and thus offers the best generalization among monocular depth estimation models. We present exemplary qualitative results of depth predictions in Fig. 7. The results reveal that strong specular reflections of the sun on the sea can disrupt the depth predictions. In the bottom two images, the camera angle is favorable, and the reflection level is low, resulting in accurate depth predictions around the water that resemble a planar surface. In contrast, the top images with reflections display clear deviations from a plane in the lower central parts of the image. Additionally, the reflection not only causes distortions in the depth predictions of the water area but also affects other objects, such as the

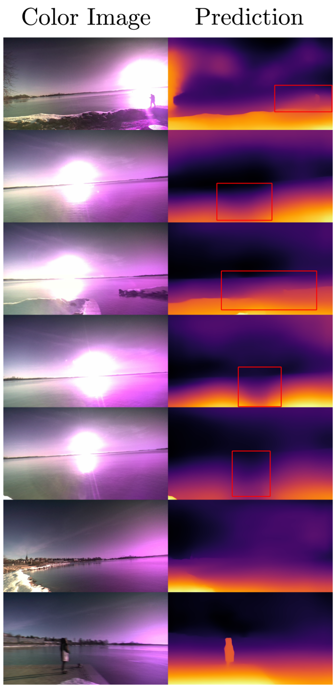


FIGURE 7. Color images from SPINS on the left and depth predictions from MIDAS on the right. The erroneous areas have been marked with red rectangles.

pedestrian in the first image and the land bridges in the third image.

C. VISUAL SLAM

This section discusses the influence of non-Lambertian surfaces on the performance of visual SLAM using TartanVO [91] and ORB SLAM 3 [93]. The established ORB SLAM 3 framework was selected for evaluation due to its state-of-the-art performance across various benchmarks. It should be noted that ORB SLAM 3 struggles with tracking and re-initialization in the monocular mode on the SPINS dataset, and thus we do not include any results from that configuration. Instead, we present the results obtained in RGB-D mode utilizing the iPhone LiDAR depth map. In contrast, we also assessed the monocular-based deep learning visual odometry (VO) model, TartanVO. It differs from ORB SLAM 3 by not maintaining an environment map and thus not susceptible to tracking loss. Additionally, as it utilizes deep learning, it can highlight possible challenges when using neural networks.

For both frameworks, a particular scene midway through the sequence poses a significant challenge for both algorithms due to strong rotational dynamics in environments characterized by specular reflections and feature sparsity. For ORB SLAM 3, these environments result in a too small amount of matched features which brings the tracking to a halt and starts again from the starting position. To maintain continuity of results for ORB SLAM, we also show results on two distinct sections of 139 m and 272 m, respectively, divided at the point of tracking failure. We computed the APE for the entire trajectory and the subdivided sections to quantify the position discrepancies at every timestep.

TABLE 4. A Table About the Results on the SPINS Dataset Using the ORB SLAM 3

Trajectory Part	Method	APE (m)
1	ORB SLAM 3	2.8604
2	ORB SLAM 3	31.3108
1	TartanVO	18.0587
2	TartanVO	76.1433
Full	TartanVO	141.8353

Table 4 presents the quantitative results, and Fig. 8 illustrates the trajectory. The first part of the trajectory (Fig. 8(a)) reveals challenges during initialization and tracking at the sequence's beginning, where the field of view is predominantly water and the motion mostly rotational. As a result, ORB SLAM 3's estimated trajectory only commences later when the camera is primarily moving forward. Despite following the correct path during the latter part of the sequence, the system does fall short of the actual position, resulting in an APE of approximately 2.8 m. This value does not account for the starting part of the sequence. TartanVO, on the other hand, accumulated a significantly larger error (2.8 m vs 18 m APE), which is due to drift into the Up axis due to to erroneous pitch estimations.

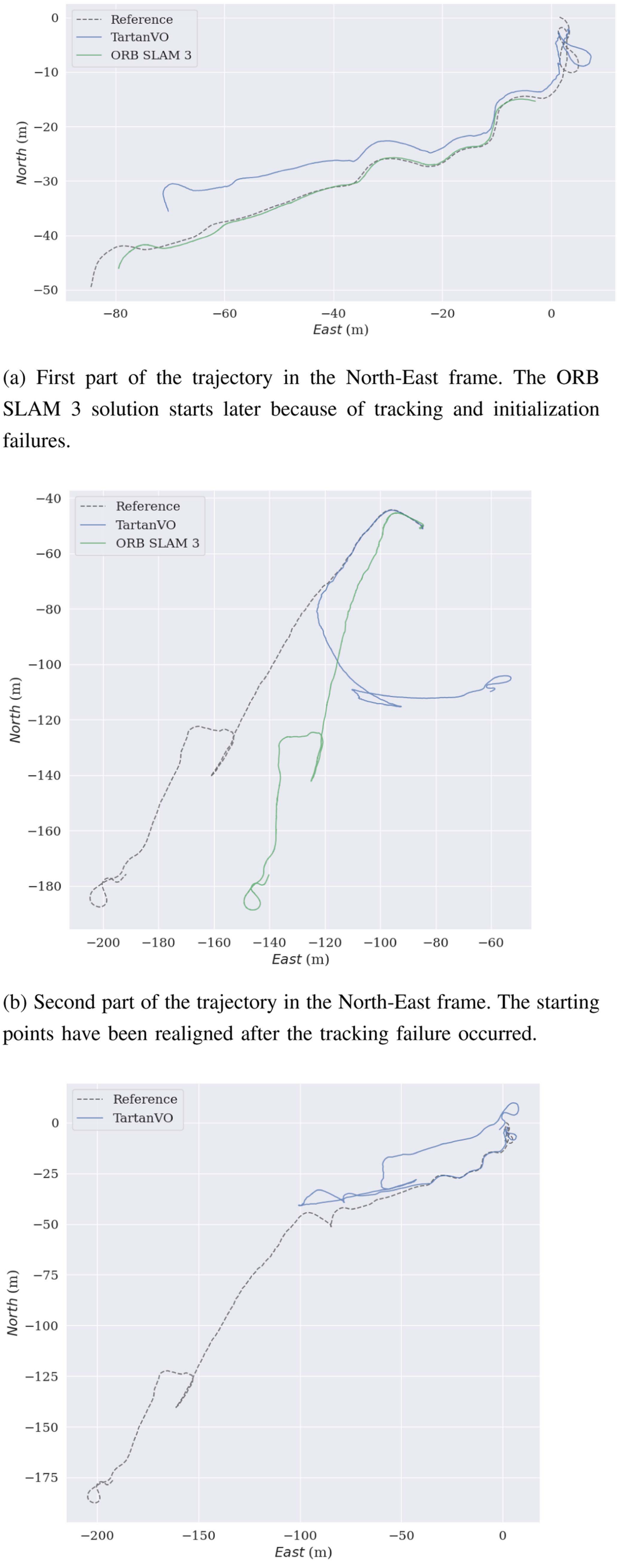
The analysis of the second trajectory section, as depicted in Fig. 8(b), reveals a substantial APE, reaching 31.3 m and 76.1 m. This increase in error is predominantly due to initial rotational estimation inaccuracies caused by an overall low number of features and erroneous feature matches. Furthermore, Fig. 8(c) demonstrates that the same region causing ORB SLAM to fail also leads to significant deviations for TartanVO, resulting in a drift away from the true path.

These findings highlight the challenges that current SLAM techniques face in environments with strong specular reflections, evident in the considerable errors from monocular methods and the tracking interruptions observed in RGB-D methods at the start and break-point of the sequence. Even after re-aligning the trajectory's second half, the positional errors significantly outpace those recorded in well-established benchmarks, such as the KITTI dataset [9], [67], emphasizing the need for enhanced robustness in SLAM algorithms under such conditions.

VI. CONCLUSION

As of now, we have gone through a wide range of studies regarding non-Lambertian surfaces in photometric stereo and monocular navigation, summarizing the modern state-of-the-art research. However, as far as the authors are aware, there is little or no data available regarding these methods' performance in industrial navigation abundant with global, dense and strong specularities.

The absence of suitable public datasets is the main hurdle for the emergence of vital new research regarding this topic.



(a) First part of the trajectory in the North-East frame. The ORB SLAM 3 solution starts later because of tracking and initialization failures.

(b) Second part of the trajectory in the North-East frame. The starting points have been realigned after the tracking failure occurred.

(c) Full SPINS trajectory of the Reference system and TartanVO [91].

FIGURE 8. Trajectory comparison of SLAM algorithms against the reference system on the SPINS.

The new dataset, SPINS, thus presents a notable milestone, providing a viable means to test navigation systems' performance with various non-Lambertian reflections. Based on the empirical results by ResNeXt-101-WSL, ORB SLAM 3 and TartanVO, we conclude that specularities affect the performance in a non-trivial fashion. We thus recommend targeted algorithm development and more extensive performance testing for monocular depth estimation and SLAM algorithms regarding non-Lambertian surfaces, to better determine their generality in challenging industrial and outdoor environments. The results can be applied in autonomous industry operations and autonomous driving, bringing forth robust and safe solutions.

As a concluding remark, more research regarding specularities and the elemental issues underlying them — the violations of brightness and colour constancies — should be conducted. Promising research avenues include integration of different information sources and robust inference methods into monocular depth estimation and SLAM, such as uncertainty maps over observations, semantic or reflection information about objects, the surface normal and gradient estimation, and hybrid models utilizing physics and deep learning. Complementary information or tailor-made inference targeting specularities. These can help computer vision algorithms become more into versatile, reliable and robust, and thus have the potential to increase the performance against featurelessness with tolerable computational costs. Especially in the special cases where the existing deep learning methods fail, such as the concave specularities, hybrid models uniting deep learning and physical models could provide beneficial effects for performance and explainability.

ACKNOWLEDGMENT

The authors thank Konecranes for their donation to support their computer vision research.

REFERENCES

[1] R. Szeliski, "Computer vision algorithms and applications," in *Computer Vision Algorithms and Applications, Texts in Computer Science*, 2nd ed.. London, U.K.: Springer, 2021, pp. 220–231.
[2] S. Hoseini and P. Kabiri, "A novel feature-based approach for indoor monocular SLAM," *Electron.*, vol. 7, no. 11, pp. 305–320, 2018.
[3] Y. Dong et al., "A novel texture-less object oriented visual SLAM system," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 1, pp. 36–49, Jan. 2021.
[4] L. Ruotsalainen, A. Morrison, M. Mäkelä, J. Rantanen, and N. Sokolova, "Improving computer vision-based perception for collaborative indoor navigation," *IEEE Sensors J.*, vol. 22, no. 6, pp. 4816–4826, Mar. 2022.
[5] J. P. Yaacoub, H. Noura, O. Salman, and A. Chehab, "Security analysis of drones systems: Attacks, limitations, and recommendations," *Internet Things*, vol. 11, 2020, Art. no. 100218.
[6] DJI, "Dji mini2 user manual," 2021. Accessed: Sep. 16, 2023. [Online]. Available:https://dl.djicdn.com/downloads/DJI_Mini_2/20210630/DJI_Mini_2_User_Manual-EN.pdf
[7] S. Kim, M. Ra, and W.-Y. Kim, "Specular detection on glossy surface using geometric characteristics of specularity in top-view images," *Sensors*, vol. 21, no. 6, 2021, Art. no. 2079.
[8] J. Ružbarský, "The difficulty of measuring the roughness of glossy surfaces using the triangulation principle," *Appl. Sci.*, vol. 13, no. 8, 2023, Art. no. 5155.

[9] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? The KITTI vision benchmark suite,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3354–3361.

[10] B. Shi, Z. Mo, Z. Wu, D. Duan, S.-K. Yeung, and P. Tan, “A benchmark dataset and evaluation for non-Lambertian and uncalibrated photometric stereo,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 271–284, Feb. 2019.

[11] Pyykölä, “Non-Lambertian surfaces and their challenges in computer vision,” Master’s thesis, Univ. Helsinki, Fac. Science, Dept. Comput. Sci., 2022.

[12] C. Cadena et al., “Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age,” *IEEE Trans. Robot.*, vol. 32, no. 6, pp. 1309–1332, Dec. 2016.

[13] W. R. McCluney, “Introduction to radiometry and photometry,” in *The Artech House Optoelectronics Library*, 2nd ed. Boston, MA, USA: Artech House, 2014.

[14] A. Earp, G. Smith, and J. Franklin, “Simplified BRDF of a non-Lambertian diffuse surface,” *Lighting Res. technol.*, vol. 39, no. 3, pp. 265–281, 2007.

[15] S. M. Rusinkiewicz, “A New Change of Variables for Efficient BRDF Representation,” in *Rendering Techniques*, G. Drettakis and N. Max, eds., Vienna, Austria: Springer, 1998, pp. 11–22.

[16] G. J. Ward, “Measuring and modeling anisotropic reflection,” in *Proc. 19th Annu. Conf. Comput. Graph. Interactive Techn.*, 1992, pp. 265–272.

[17] G. Zhang and P. A. Vela, “Good features to track for visual SLAM,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1373–1382.

[18] O. Meslouhi, M. Kardouchi, H. Allali, T. Gadi, and Y. Benkaddour, “Automatic detection and inpainting of specular reflections for colposcopic images,” *Open Comput. Sci.*, vol. 1, no. 3, pp. 341–354, 2011.

[19] A. Mirko, G. Anarta, A. Stefan, and G. Lacey, “Automatic segmentation and inpainting of specular highlights for endoscopic imaging,” *EURASIP J. Image Video Process.*, vol. 2010, pp. 1–12, 2010.

[20] C.-A. Saint-Pierre, J. Boisvert, G. Grimard, and F. A. Cheriet, “Detection and correction of specular reflections for automatic surgical tool segmentation in thoracoscopic images,” *Mach. Vis. Appl.*, vol. 22, no. 1, pp. 171–180, 2007.

[21] R. J. Woodham, “Photometric method for determining surface orientation from multiple images,” *Opt. Eng.*, vol. 19, no. 1, pp. 139–144, 1980.

[22] J. P. David and A. Forsyth, *Computer Vision: A Modern Approach*, 2nd ed. Englewood Cliffs, NJ, USA: Prentice Hall, 2011.

[23] T.-P. Wu and C.-K. Tang, “Photometric stereo via expectation maximization,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 3, pp. 546–560, Mar. 2010.

[24] E. N. Coleman and R. Jain, “Obtaining 3-dimensional shape of textured and specular surfaces using four-source photometry,” *Comput. Graph. Image Process.*, vol. 18, no. 4, pp. 309–328, 1982.

[25] F. Verbiest and L. Van Gool, “Photometric stereo with coherent outlier handling and confidence estimation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[26] S. Ikehata, D. Wipf, Y. Matsushita, and K. Aizawa, “Robust photometric stereo using sparse regression,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 318–325.

[27] M. Pharr, W. Jakob, and G. Humphreys, *Physically Based Rendering: From Theory To Implementation*, 3rd ed. Cambridge, MA, USA: MIT Press, 2018. Accessed: Sep. 5, 2022. [Online]. Available: https://www.pbr-book.org/3ed-2018/contents

[28] Georghiades, “Incorporating the torrance and sparrow model of reflectance in uncalibrated photometric stereo,” in *Proc. IEEE 9th Int. Conf. Comput. Vis.*, 2003, vol. 2, pp. 816–823.

[29] H.-S. Chung and J. Jia, “Efficient photometric stereo on glossy surfaces with wide specular lobes,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[30] D. B. Goldman, B. Curless, A. Hertzmann, and S. M. Seitz, “Shape and spatially-varying BRDFs from photometric stereo,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 6, pp. 1060–1071, Jun. 2010.

[31] J. Ackermann, F. Langguth, S. Fuhrmann, and M. Goesele, “Photometric stereo for outdoor webcams,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 262–269.

[32] X. Wang, Z. Jian, and M. Ren, “Non-Lambertian photometric stereo network based on inverse reflectance model with collocated light,” *IEEE Trans. Image Process.*, vol. 29, pp. 6032–6042, 2020.

[33] B. Shi, P. Tan, Y. Matsushita, and K. Ikeuchi, “Bi-polynomial modeling of low-frequency reflectances,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 6, pp. 1078–1091, Jun. 2014.

[34] J. Lim, J. Ho, M.-H. Yang, and D. Kriegman, “Passive photometric stereo from motion,” in *Proc. IEEE 10th Int. Conf. Comput. Vis. Volume 1*, 2005, vol. 2, pp. 1635–1642.

[35] Simakov, Frolova and Basri, “Dense shape reconstruction of a moving object under arbitrary, unknown lighting,” in *Proc. IEEE 9th Int. Conf. Comput. Vis.*, 2003, vol. 2, pp. 1202–1209.

[36] M. Zhou, Y. Ding, Y. Ji, S. S. Young, J. Yu, and J. Ye, “Shape and reflectance reconstruction using concentric multi-spectral light field,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 7, pp. 1594–1605, Jul. 2020.

[37] S. M. Haque, A. Chatterjee, and V. M. Govindu, “High quality photometric reconstruction using a depth camera,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2283–2290.

[38] Q. Zhang, M. Ye, R. Yang, Y. Matsushita, B. Wilburn, and H. Yu, “Edge-preserving photometric stereo via depth fusion,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2472–2479.

[39] G. Chen, K. Han, B. Shi, Y. Matsushita, and K.-Y. K. Wong, “Deep photometric stereo for non-Lambertian surfaces,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 129–142, Jan. 2022.

[40] B. Kaya, S. Kumar, C. Oliveira, V. Ferrari, and L. Van Gool, “Uncalibrated neural inverse rendering for photometric stereo of general surfaces,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 3804–3814.

[41] T. Taniai and T. Maehara, “Neural inverse rendering for general reflectance photometric stereo,” in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 4857–4866.

[42] S. Georgoulis et al., “Reflectance and natural illumination from single-material specular objects using deep learning,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 8, pp. 1932–1947, Aug. 2018.

[43] X. Li, Y. Dong, P. Peers, and X. Tong, “Modeling surface appearance from a single photograph using self-augmented convolutional neural networks,” *ACM Trans. Graph.*, vol. 36, pp. 1–11, 2017.

[44] C.-H. Rhee and C. H. Lee, “Estimating physically-based reflectance parameters from a single image with GAN-guided CNN,” *IEEE Access*, vol. 10, pp. 13259–13269, 2022.

[45] L. Sang, B. Häfner, X. Zuo, and D. Cremers, “High-quality RGB-D reconstruction via multi-view uncalibrated photometric stereo and gradient-SDF,” in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Jan. 2023, pp. 3106–3115.

[46] B. Shi, Y. Matsushita, Y. Wei, C. Xu, and P. Tan, “Self-calibrating photometric stereo,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 1118–1125.

[47] T. Papadhimitri and P. Favaro, “A closed-form, consistent and robust solution to uncalibrated photometric stereo via local diffuse reflectance maxima,” *Int. J. Comput. Vis.*, vol. 107, pp. 139–154, 2014.

[48] N. G. Alldrin, S. P. Mallick, and D. J. Kriegman, “Resolving the generalized bas-relief ambiguity by entropy minimization,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–7.

[49] G. Chen, K. Han, B. Shi, Y. Matsushita, and K.-Y. K. K. Wong, “Self-calibrating deep photometric stereo networks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8731–8739.

[50] I. Sato, T. Okabe, Q. Yu, and Y. Sato, “Shape reconstruction based on similarity in radiance changes under varying illumination,” in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, 2007, pp. 1–8.

[51] F. Lu, I. Sato, and Y. Sato, “Uncalibrated photometric stereo based on elevation angle recovery from BRDF symmetry of isotropic materials,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 168–176.

[52] K. Yamashita, Y. Enyo, S. Nobuhara, and K. Nishino, “nLMVS-Net: Deep non-Lambertian multi-view stereo,” in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2023, pp. 3037–3046.

[53] J. Lv, H. Guo, G. Chen, J. Liang, and B. Shi, “NeuraIMPS: Non-Lambertian multispectral photometric stereo via spectral reflectance decomposition,” 2022, *arXiv:2211.15311*.

[54] D. Chen, P. Zhang, I. Feldmann, O. Schreer, and P. Eisert, “Recovering fine details for neural implicit surface reconstruction,” in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2023, pp. 4330–4339.

[55] T. Iwaguchi and H. Kawasaki, “Surface normal estimation from optimized and distributed light sources using DNN-based photometric stereo,” in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2023, pp. 311–320.

[56] H. Hayakawa, “Photometric stereo under a light source with arbitrary motion,” *J. Opt. Soc. Amer. A-Opt. Image Sci. Vis.*, vol. 11, pp. 3079–3089, 1994.

[57] A. Del Bue, J. Xavier, L. Agapito, and M. Paladini, “Bilinear factorization via augmented Lagrange multipliers,” in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 283–296.

[58] L. Ruotsalainen, M. Kirkko-Jaakkola, and J. Talvitie, *The Present and Future of Indoor Navigation*. Norwood, MA, USA: Artech House, Nov. 2023.

[59] H. Durrant-Whyte and T. Bailey, “Simultaneous localization and mapping: Part I,” *IEEE Robot. Automat. Mag.*, vol. 13, no. 2, pp. 99–110, Jun. 2006.

[60] J. D. Paul, W. Buytaert, and N. Sah, “A technical evaluation of LiDAR-based measurement of river water levels,” *Water Resour. Res.*, vol. 56, no. 4, 2020, Art. no. e2019WR026810, doi: 10.1029/2019WR026810.

[61] V. Ila, L. Polok, M. Solony, and P. Svoboda, “SLAM++ -A highly efficient and temporally scalable incremental SLAM framework,” *Int. J. Robot. Res.*, vol. 36, 2017, Art. no. 027836491769111.

[62] R. A. Newcombe et al., “KinectFusion: Real-time dense surface mapping and tracking,” in *Proc. 10th IEEE Int. Symp. Mixed Augmented Reality*, 2011, pp. 127–136.

[63] F. Chhaya, D. Reddy, S. Upadhyay, V. Chari, M. Z. Zia, and K. M. Krishna, “Monocular reconstruction of vehicles: Combining SLAM with shape priors,” in *Proc. IEEE Int. Conf. Robot. Automat.*, 2016, pp. 5758–5765.

[64] S. Yang and S. Scherer, “CubeSLAM: Monocular 3-D object SLAM,” *IEEE Trans. Robot.*, vol. 35, no. 4, pp. 925–938, Aug. 2019.

[65] C. Häne, N. Savinov, and M. Pollefeys, “Class specific 3D object shape priors using surface normals,” in *Proc. 2014 IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 652–659.

[66] A. Dame, V. A. Prisacariu, C. Y. Ren, and I. Reid, “Dense reconstruction using 3D object shape priors,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 1288–1295.

[67] R. Mur-Artal and J. D. Tardós, “ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras,” *IEEE Trans. Robot.*, vol. 33, no. 5, pp. 1255–1262, Oct. 2017.

[68] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, “ORB: An efficient alternative to SIFT or SURF,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 2564–2571.

[69] K. Tateno, F. Tombari, I. Laina, and N. Navab, “CNN-SLAM: Real-time dense monocular SLAM with learned depth prediction,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 2017, pp. 6243–6252.

[70] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[71] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, “DeepSDF: Learning continuous signed distance functions for shape representation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 165–174.

[72] E. Sucar, K. Wada, and A. Davison, “NodeSLAM: Neural object descriptors for multi-view shape reconstruction,” in *Proc. Int. Conf. 3D Vis.*, 2020, pp. 949–958.

[73] K. Li et al., “FroDO: From detections to 3D objects,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 14708–14717.

[74] N. Yang, L. von Stumberg, R. Wang, and D. Cremers, “D3VO: Deep depth, deep pose and deep uncertainty for monocular visual odometry,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1278–1289.

[75] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[76] J. Wang, M. Rünz, and L. Agapito, “DSP-SLAM: Object oriented SLAM with deep shape priors,” in *Proc. IEEE Int. Conf. 3D Vis.*, 2021, pp. 1362–1371.

[77] J. Engel, V. Koltun, and D. Cremers, “Direct sparse odometry,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 3, pp. 611–625, Mar. 2018.

[78] B. Bescos, J. M. Facil, J. Civera, and J. Neira, “DynaSLAM: Tracking, mapping, and inpainting in dynamic scenes,” *IEEE Robot. Automat. Lett.*, vol. 3, no. 4, pp. 4076–4083, Oct. 2018.

[79] L. Xiao, J. Wang, X. Qiu, Z. Rong, and X. Zou, “Dynamic-SLAM: Semantic monocular visual localization and mapping based on deep learning in dynamic environment,” *Robot. Auton. Syst.*, vol. 117, pp. 1–16, 2019.

[80] J. Mo, M. J. Islam, and J. Sattar, “Fast direct stereo visual SLAM,” *IEEE Robot. Automat. Lett.*, vol. 7, no. 2, pp. 778–785, Apr. 2022.

[81] D. Antensteiner, S. Štolc, and T. Pock, “A review of depth and normal fusion algorithms,” *Sensors*, vol. 18, 2018, Art. no. 431.

[82] H. Zhan, C. S. Weerasekera, R. Garg, and I. Reid, “Self-supervised learning for single view depth and surface normal estimation,” in *Proc. IEEE Int Conf. Robot. Automat.*, 2019, pp. 4811–4817.

[83] L. Zhang, Hertzmann, Curless, and Seitz, “Shape and motion under varying illumination: Unifying structure from motion, photometric stereo, and multiview stereo,” in *Proc. IEEE 9th Int. Conf. Comput. Vis.*, 2003, vol. 1, pp. 618–625.

[84] N. Joshi and D. J. Kriegman, “Shape from varying illumination and viewpoint,” in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, 2007, pp. 1–7.

[85] D. Slater and G. Healey, “The illumination-invariant recognition of 3D objects using local color invariants,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 2, pp. 206–210, Feb. 1996.

[86] M. Labbé and F. Michaud, “Multi-session visual SLAM for illumination-invariant re-localization in indoor environments,” *Front. Robot. AI*, vol. 9, 2022, Art. no. 801886.

[87] X. Zou, J. Kittler, and K. Messer, “Illumination invariant face recognition: A survey,” in *Proc. IEEE 1st Int. Conf. Biometrics: Theory, Appl., Syst.*, 2007, pp. 1–8.

[88] J. Engel, V. Usenko, and D. Cremers, “A photometrically calibrated benchmark for monocular visual odometry,” 2016, *arXiv:1607.02555*.

[89] J.-L. Blanco-Claraco, F. Á. Moreno-Dueñas, and J. González-Jiménez, “The Málaga urban dataset: High-rate stereo and LiDAR in a realistic urban scenario,” *Int. J. Robot. Res.*, vol. 33, no. 2, pp. 207–214, 2014.

[90] S. Umeyama, “Least-squares estimation of transformation parameters between two point patterns,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 4, pp. 376–380, Apr. 1991.

[91] W. Wang, Y. Hu, and S. Scherer, “TartanVO: A generalizable learning-based VO,” in *Proc. Conf. Robot Learn.*, 2021, pp. 1761–1772.

[92] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, “Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1623–1637, Mar. 2022.

[93] C. Campos, R. Elvira, J. J. G. Rodriguez, J. M. M. Montiel, and J. D. Tardós, “ORB-SLAM3: An accurate open-source library for visual, visual-inertial, and multimap SLAM,” *IEEE Trans. Robot.*, vol. 37, no. 6, pp. 1874–1890, Dec. 2021.

SARA PYYKÖLÄ is currently working toward the Ph.D. degree in computer science with the University of Helsinki, Helsinki, Finland. Her research interests include multi-objective reinforcement learning, computer vision, and sustainable AI.

NICLAS JOSWIG is currently working toward the Ph.D. degree in computer science with the University of Helsinki, Helsinki, Finland. His research interests include computer vision, navigation, and autonomous machines.

LAURA RUOTSALAINEN is currently a Professor of spatiotemporal data analysis for sustainability science with the Department of Computer Science, University of Helsinki, Helsinki, Finland. She leads a research group which performs research on machine learning and computer vision using spatiotemporal data for sustainable smart cities especially via smart mobility. She is a Member of the Steering Group of the Finnish Center for AI (FCAI) and lead in AI and Sustainability. She is also with the Helsinki Institute of Sustainability Science (HELSUS).