Master's thesis
Master's Programme in Data Science

Non-Lambertian surfaces and their challenges in computer vision

Sara Pyykölä

September 6, 2022

Supervisor(s): Assoc. Prof. Laura Ruotsalainen
MSc Niclas Joswig

Examiner(s): Assoc. Prof. Laura Ruotsalainen
Prof. Jukka K.Nurminen

University of Helsinki
Faculty of Science

P. O. Box 68 (Pietari Kalmin katu 5)
00014 University of Helsinki

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Koulutusohjelma — Utbildningsprogram — Degree programme
Faculty of Science	Master's Programme in Data Science

Tekijä — Författare — Author

Sara Pyykölä

Työn nimi — Arbetets titel — Title

Non-Lambertian surfaces and their challenges in computer vision

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidantal — Number of pages
Master's thesis	September 6, 2022	64

Tiivistelmä — Referat — Abstract

This thesis regards non-Lambertian surfaces and their challenges, solutions and study in computer vision. The physical theory for understanding the phenomenon is built first, using the Lambertian reflectance model, which defines Lambertian surfaces as ideally diffuse surfaces, whose luminance is isotropic and the luminous intensity obeys Lambert's cosine law. From these two assumptions, non-Lambertian surfaces violate at least the cosine law and are consequently specularly reflecting surfaces, whose perceived brightness is dependent from the viewpoint. Thus non-Lambertian surfaces violate also brightness and colour constancies, which assume that the brightness and colour of same real-world points stays constant across images. These assumptions are used, for example, in tracking and feature matching and thus non-Lambertian surfaces pose complications for object reconstruction and navigation among other tasks in the field of computer vision.

After formulating the theoretical foundation of necessary physics and a more general reflectance model called the bi-directional reflectance distribution function, a comprehensive literature review into significant studies regarding non-Lambertian surfaces is conducted. The primary topics of the survey include photometric stereo and navigation systems, while considering other potential fields, such as fusion methods and illumination invariance. The goal of the survey is to formulate a detailed and in-depth answer to what methods can be used to solve the challenges posed by non-Lambertian surfaces, what are these methods' strengths and weaknesses, what are the used datasets and what remains to be answered by further research. After the survey, a dataset is collected and presented, and an outline of another dataset to be published in an upcoming paper is presented. Then a general discussion about the survey and the study is undertaken and conclusions along with proposed future steps are introduced.

ACM Computing Classification System (CCS):
Computing methodologies → Artificial intelligence → Computer vision → Computer vision problems

Avainsanat — Nyckelord — Keywords

Non-Lambertian surfaces, photometric stereo, SLAM, monocular depth estimation

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — Övriga uppgifter — Additional information

Contents

Introduction

Computer vision has since 1970s sought the visionary ideal of a machine that could see and perceive like a human being [58]. In spite of the blooming optimism of the early days and the splendour consequences of the realization of that ideal, the ultimate solutions have turned out to be elusive for a myriad of reasons and the research on computer vision remains versatile and lively to this day, taking on challenges ranging from cost-effective imaging to the accurate reconstructions of objects and flawless image matching, as extensively introduced by Szeliski in his book "Computer Vision Algorithms and Applications" from 2021 [58].

One of the essential challenges of modern computer vision lies in what I shall refer to as *featurelessness*, which can be roughly translated as the lack of anything to look at or to be observed by image sensors. A few practical examples include uniformly coloured settings and strongly reflecting surfaces, the latter referred to as *non-Lambertian surfaces* [25]. While the immediately obvious consequence is the absence of useful quantifiable information, one must consider also their secondary side effect, which disturbs the functionality of algorithms as a whole. Non-Lambertian surfaces have proven to be quite the hurdle, for example, in medical endoscopies [50], where the closed environment of human body permits no changes, the mucous body fluids and metallic tools produce specularities, and the primary equipment is an endoscope unable to perform major adjustments to itself after the start of operation. It has thus led to sophisticated masking algorithms capable of assisting medical professionals to see vital details in real-time during the operations, despite the specularities present [53].

Naturally these bright spotlights have been spotted in other fields of computer vision as well, such as navigation [16]. As the cameras and algorithms can not, metaphorically speaking, put a finger on it what they're observing, the tasks of localization and image patch feature matching become increasingly difficult for the regions occupied with the light [25]. It thus brings forth the notion of significant loss in accuracy due to this phenomenon and the inevitable research question of how to effectively address the issue in terms of the algorithm and equipment.

In this thesis, I shall formulate answers to three questions: "what methods can be used to extract useful information from non-Lambertian surfaces for computer vision algorithms in different fields', "how good exactly are those methods" and "what kind of datasets are used". This is done by the means of a literature survey, where from the reader I assume the basic knowledge of computer vision, such as the definitions

of camera calibration and projection matrices, familiarity with features and feature matching algorithms and a basic understanding of monocular cameras. Necessary knowledge can be gained, for example, from chapters 2, 7 and 10 in "Computer Vision: Algorithms and Applications" by Szeliski [58], or chapters 1, 5 and 7 in "Computer Vision: Modern Approach" by Forsyth and Ponce [13].

I will then start by presenting the fundamental theoretical foundation and physical models of non-Lambertian surfaces, and then continue on to present relevant studies from various fields of computer vision. After the survey, I'll present the commonly used datasets in the surveyed methods, collect and test a new dataset aimed at quantifying the non-Lambertian surfaces' exact effect for single-view monocular depth estimation and close off by proposing another dataset regarding tracking and mapping, to be published in a paper in near future. Thirdly a discussion about pros and cons of the conducted empirical studies takes place, in which I shall suggest ideas for further research by regarding the existing state-of-the art methods and the results of the conducted empirical studies. The thesis is closed by conclusions from the survey and the future steps. I argue the benefits of such survey, studies and discussions lie in the extensive viewpoint from which non-Lambertian surfaces are studied on across vastly different fields of computer vision. Such approach is to my knowledge first of its kind and thus it may instigate fresh research avenues for the upcoming state-of-the-art methods aiming to solve the issue posed by non-Lambertian surfaces.

1. 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 [5, 53]. Here I have adopted the problem-centric approach, where the problem and its physical models are presented first and its solutions afterwards. I claim that by understanding the problem and phenomenon itself in great detail, the solutions become consequently clearer to the reader, and moreover the scope of the literature survey can be limited to the most relevant parts regarding the theoretical foundation of proposed future research. Thus, this chapter considers the surface reflectance's physical models and the definitions of non-Lambertian surfaces as formulated by the models. If the reader is acquainted with elemental physics regarding light and energy, they may freely skip to following section and refer to Table 1.1 for exact units and Table 1.2 for descriptions of quantities, as needed. Likewise readers further acquainted with the Lambertian reflectance model and bi-directional reflection distribution function may skip to the following chapter, referring to Equations (1.2) and (1.1), and Tables 1.1 and 1.2 as needed.

1.1 Surface reflectance models

1.1.1 Surface reflectance

To understand surface reflectance models, one must first define what exactly is *surface reflectance*. Simply put, surface reflectance is defined as light the surface reflects when light is emitted at it, as explained in McCluney's book "Introduction to radiometry and photometry" (2014) [37]. Continuing from this generally accepted viewpoint, the light can be physically perceived in three major ways: either as electromagnetic radiation in radiometry or more directly as the human's sensory perception of brightness in photometry or as the behaviour of a specific energy source in optics. On account of these differences in focus, McCluneuy demonstrated there are a few distinctions between the fields, such as the photometry's limited scope of only visible light opposed to other two, but the basic quantities and units of measurement, which enable us to quantify light for the surface reflectance models, are still largely the same. Assuming only an elemental notion of geometry and waves — namely, the concepts of wavelength, frequency and radian — from the reader, I'll present a short introduction into these quantities and units next and the surface reflectance models in the next section. For a

General concept	**Quantity**	
Name	Name	Description
Flux	Radiant flux	Radiant energy per unit time
	Spectral flux	Radiant flux per unit frequency or wavelength
	Luminous flux	Luminous energy per time unit
Intensity	Radiant intensity	Radiant flux per unit angle
	Spectral intensity	Radiant intensity per unit frequency or wavelength
	Luminuous intensity	Luminous flux per unit angle
Radiance	Radiance	Radiant flux by a surface, per unit angle per unit area
	Spectral radiance	Radiance per unit frequency or wavelength
Irradiance	Irradiance	Radiant flux received by a surface per unit area
	Spectral irradiance	Irradiance per unit frequency or wavelength
Exitance	Radiant exitance	Radiant flux emitted by a surface per unit area
	Spectral exitance	Radiant exitance per unit frequency or wavelength
	Luminous exitance	Luminous flux per unit angle per unit projected source area
Radiosity	Radiosity	Radiant flux leaving a surface per unit area
	Spectral radiosity	Radiosity of a surface per unit frequency or wavelength
Luminance	Luminance	Luminous flux per unit solid angle per unit projected source area
	Illuminance	Luminous flux incident on a surface
Surface albedo	Surface albedo	Ratio of radiosity to the irradiance

Table 1.1: A table describing the different physical quantities related to light.

more detailed overview, I'll refer the reader to the aforementioned McCluney's book, to which the rest of this section is based on [37].

Three central quantities of radiant energy include radiant intensity, radiance and irradiance, which are all constructively founded on another central quantity, radiant flux [37]. Radiant flux is the radiant energy transferred per time unit, measured as watts (denoted as W). Radiant intensity is then emitted radiant flux per unit angle, and is measured as watts per steradian ($\mathrm{W} \cdot \mathrm{sr}^{-1}$), where steradian is a three-dimensional unit angle, analogous to two-dimensional radian. Respectively radiance is radiant flux by a surface, per unit angle per unit area and is measured watts per steradian per square metre ($\mathrm{W} \cdot \mathrm{sr}^{-1} \cdot \mathrm{m}^{-2}$). Finally, irradiance is radiant flux received by a surface area and is measured as watts per square metre ($\mathrm{W} \cdot \mathrm{m}^{-2}$). Here it is important to note the small difference between radiance and irradiance: the former refers to any transferred light with respect to a surface, whether it was reflected, transmitted, received or emitted, while the latter only concerns the light received by the surface. Thus the unit is different as well.

All of these quantities have also their distinct equivalents in the context of electromagnetic radiation [37]. They are called spectral quantities, namely spectral flux, spectral intensity, spectral radiance and spectral irradiance. Their difference to the radiant quantities is minute, but vital: all is measured as per unit frequency or wavelength. Consequently, the units of measurement depend on whether the unit frequency (Hz) or wavelength (metres) is being used. Spectral flux is then measured as radiant flux per time unit per wavelength or unit frequency, and the unit can be either watts per hertz or watts per metre. A similar reasoning can be followed for the rest: spectral intensity is radiant intensity per wavelength or unit frequency, spectral radiance is radiance per wavelength or unit frequency and spectral irradiance is irradiance per wavelength or unit frequency [37]. For the exact units of measurement, I refer the reader to Table 1.1; for descriptions regarding the quantities, I refer to Table 1.2.

In addition to the aforementioned quantities, photometry has a few equivalent quantities of its own: luminous flux, luminous intensity, luminance and illuminance. They are roughly equivalent to their radiant and spectral counterparts in idea, but the light waves are weighted according to a luminosity function, leading to different units. The unit for luminous flux is lumen, equal to candela steradian, measuring the luminous energy per time unit. The rest of the luminant units follow the logic of radiant and spectral units, but are measured in terms of lumens. The Table 1.1 contains the exact units for these quantities, as well as for radiant and spectral radiosity and radiant, spectral and luminous exitance, which respectively distinct radiant flux exiting a surface or emitted by a surface. The descriptions for all aforementioned quantities can be found in Table 1.2. As a final remark, a surface albedo is a commonly used

General concept	**Quantity**		**Unit**
Name	Name	Symbol	Symbol (name)
Flux	Radiant flux	Φ_e	$\mathrm{W} \cdot \mathrm{sr}^{-1}$
	Spectral flux	$\Phi_{e,\lambda}$ or $\Phi_{e,v}$	$\mathrm{W} \cdot \mathrm{sr}^{-1}$
	Luminous flux	$\Phi_{e,\lambda}$ or $\Phi_{e,v}$	$\mathrm{W} \cdot \mathrm{sr}^{-1}$
Intensity	Radiant intensity	$I_{e,\Omega}$	$\mathrm{W} \cdot \mathrm{sr}^{-1}$
	Spectral intensity	$I_{e,v,\lambda}$	$\mathrm{W} \cdot \mathrm{sr}^{-1} \cdot \mathrm{m}^{-1}$ or $\mathrm{W} \cdot \mathrm{sr}^{-1} \cdot \mathrm{Hz}^{-1}$
	Luminuous intensity	I_v	$\mathrm{cd} = \mathrm{lm} \cdot \mathrm{sr}^{-1}$ (candela)
Radiance	Radiance	$L_{e,\Omega}$	$\mathrm{W} \cdot \mathrm{sr}^{-1} \cdot \mathrm{m}^{-2}$
	Spectral radiance	$L_{e,\Omega,\lambda}$	$\mathrm{W} \cdot \mathrm{sr}^{-1} \cdot \mathrm{m}^{-2} \cdot \mathrm{Hz}^{-1}$ or $\mathrm{W} \cdot \mathrm{sr}^{-1} \cdot \mathrm{m}^{-3}$
Irradiance	Irradiance	E_e	$\mathrm{W} \cdot \mathrm{m}^{-1}$
	Spectral irradiance	$E_{e,v}$ or $E_{e,\lambda}$	$\mathrm{W} \cdot \mathrm{m}^{-2} \cdot \mathrm{Hz}^{-1}$ or $\mathrm{W} \cdot \mathrm{m}^{-3}$
Exitance	Radiant exitance	M_e	$\mathrm{W} \cdot \mathrm{m}^{-2}$
	Spectral exitance	$M_{e,v}$ or $M_{e,\lambda}$	$\mathrm{W} \cdot \mathrm{m}^{-2} \cdot \mathrm{Hz}^{-1}$ or $\mathrm{W} \cdot \mathrm{m}^{-3}$
	Luminous exitance	M_v	$\mathrm{lm} \cdot \mathrm{m}^{-2}$
Radiosity	Radiosity	J_e	$\mathrm{W} \cdot \mathrm{m}^{-2}$
	Spectral radiosity	$J_{e,v}$ or $J_{e,\lambda}$	$\mathrm{W} \cdot \mathrm{m}^{-2} \cdot \mathrm{Hz}^{-1}$ or $\mathrm{W} \cdot \mathrm{m}^{-3}$
Luminance	Luminance	L_v	$\mathrm{cd} \cdot \mathrm{m}^{-2}$
	Illuminance	E_v	$\mathrm{lx} = \mathrm{lm} \cdot \mathrm{m}^{-2}$ (lux)

Table 1.2: A table detailing units of measurement and symbols for different physical quantities related to light.

measurement in the computations to describe the surface's ability to reflect light, and is defined as the ratio of radiosity to the irradiance received by a surface [37]. However, as this is a dimensionless quantity without a symbol and unit of its own, it is omitted from Table 1.1 and presented in Table 1.2 only.

As we have now gone over the basic operationalization of light, it is time to ask how these quantities enable us to model the light. The key quantities are the luminous intensity and luminous radiance, which capture the light received by the surface in a directional fashion. Let us next present the two models based on these quantities, the first of which coined the term for non-Lambertian surfaces and the second which generalized the theory behind the first model [37]. Through out the next section, while

it may be cumbersome, I shall use the notation from Table 1.1 in quantities' subscripts to absolve any ambiguity regarding the measurement units for less physically oriented readers. Hence I advice the reader to refer to the Table 1.1, when interpreting the equations.

1.1.2 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 [18]. 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; $I_{v;0}$, the reflected luminous intensity of the surface; and the surface colour C [18]:

$$I_{v;r} = C \cdot I_{v;0} \cdot \cos\theta.$$

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 [18]. Then the model can be expressed in the vector form:

$$I_{v;r} = \underbrace{\langle \mathbf{n}, \mathbf{l} \rangle}_{\cos\theta} \cdot C \cdot I_{v;0},$$

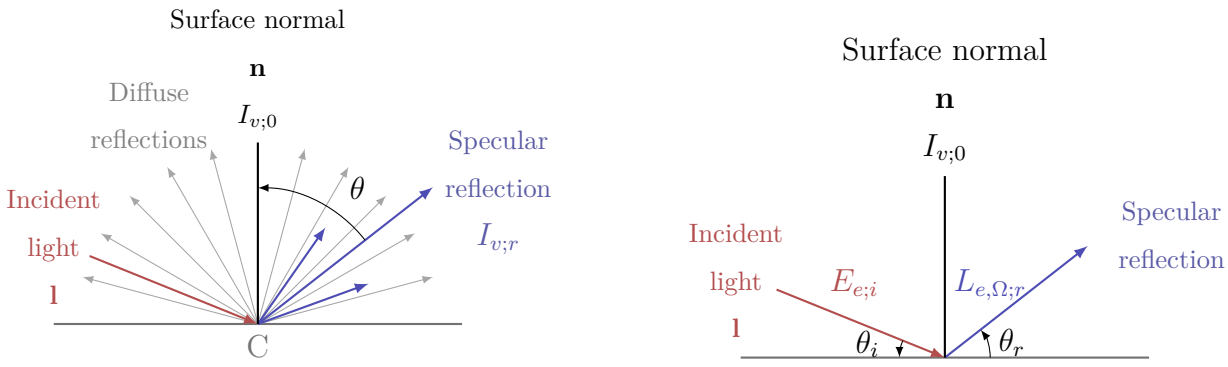
where $\langle \cdot\,, \cdot \rangle$ denotes dot product, $\mathbf{n}$ is the surface normal vector and $\mathbf{l}$ is the normalized incident light's direction vector [18].

This model has coined diffusely reflecting surfaces as Lambertian surfaces and their opposite as non-Lambertian surfaces, while the reflections are diffuse or specular correspondingly [49]. The visual difference between these two reflections is illustrated in Table 1.3; intuitively it can described as if the reflection is concentrated on a random surface point (non-Lambertian, specular reflection), instead of being scattered evenly around the point where the light hits the surface (Lambertian, diffuse reflection). 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 Figure 1.1a [18].

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 [18].



(a) The Lambertian reflectance model.

(b) The BDRF model.

Figure 1.1: 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, $I_{v;r}$ for the reflection's intensity and l for the directional vector of incident light. $E_{e;i}$ denotes the incident light's irradiance and $L_{e,\Omega;r}$ 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.

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. Let us investigate that next.

1.1.3 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 [49]. Mathematically put, it is the ratio of reflected ray's radiance, $L_{e,\Omega;\,r}$, to the incident ray's irradiance, $E_{e;\,i}$, given their directions with respect to the surface normal $\mathbf{n}$, denoted as θ_r and θ_i and called the elevation angles [49]:

$$f_r(\theta_r, \theta_i) = \frac{dL_{e,\Omega;\,r}(\theta_r)}{dE_{e;\,i}(\theta_i)}$$

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 [65], a property which is called the Helmholtz reciprocity [58].

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 [49]:

$$f_r(\theta_r, \theta_i, \phi_r, \phi_i) = \frac{dL_{e,\Omega;\,r}(\theta_r, \phi_r)}{dE_{e;\,i}(\theta_i, \phi_i)}.$$

As the rays trail along the surface of a cone centered at the surface normal's intersection point with the surface, the angles also define the said cone's radius and height. Thus, the three-dimensional BRDF can additionally be parametrized with a half-angle bisector coordinate system, formulated as the following equation,

$$f_r(\theta_d, \theta_h, \phi_d, \phi_h) = \frac{dL_{e,\Omega;\,r}(\theta_h, \phi_h)}{dE_{e;\,i}(\theta_d, \phi_d)},$$

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 [49]. The different coordinate systems are further illustrated in Figure 1.2. 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 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 three-dimensional surface normal vector,
- $\mathbf{v}$ is the three-dimensional viewing point,
- $\mathbf{h}$ is the three-dimensional half-angle vector and
- $\mathbf{l}$ is the three-dimensional incident light vector [53].

Then the BRDF becomes

$$f_r(\theta_r, \theta_i, \phi_r, \phi_i) = f_r(\mathbf{n}, \mathbf{l}), \tag{1.1}$$

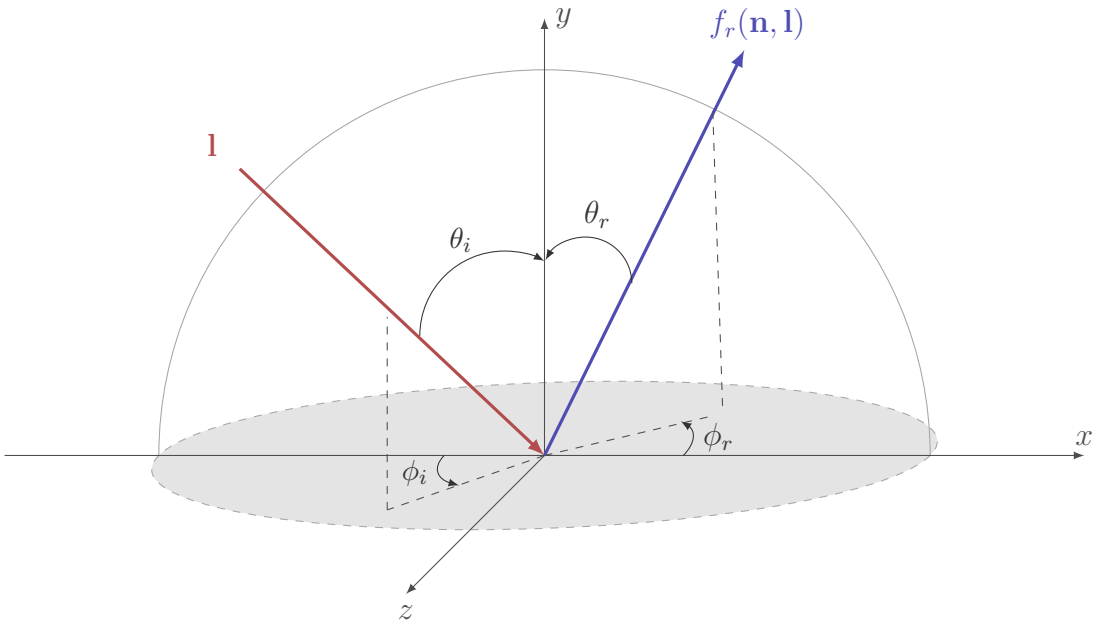
$$f_r(\theta_d, \theta_h, \phi_d, \phi_h) = f_r(\mathbf{v}, \mathbf{l}) \tag{1.2}$$

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

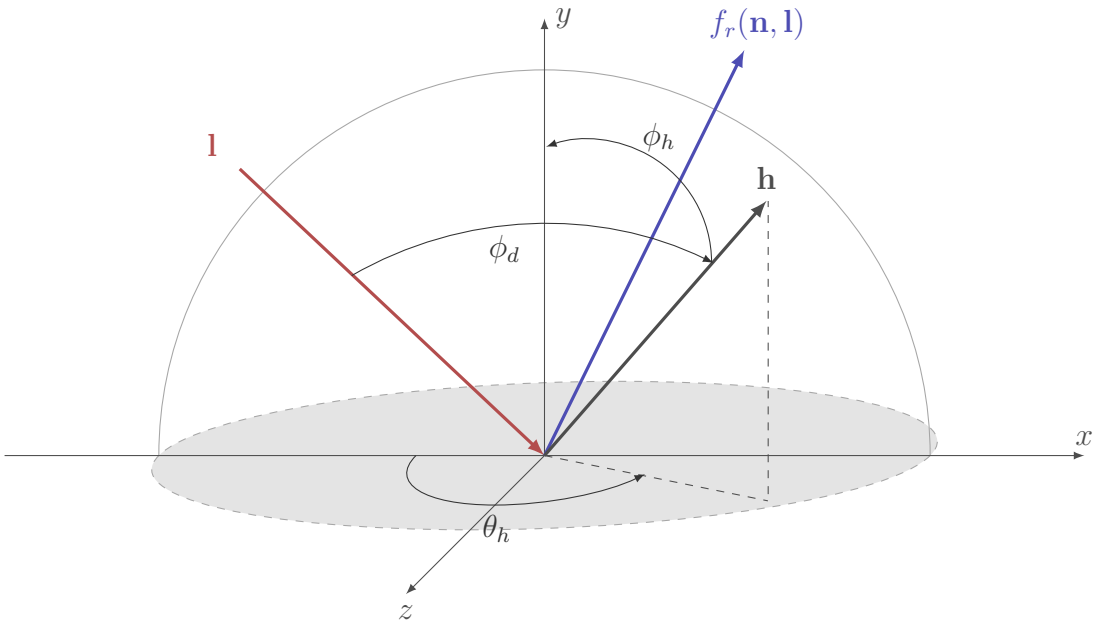
$$\theta_h = \langle \mathbf{n}, \mathbf{h} \rangle = \arccos\left(\mathbf{n}^T \mathbf{h}\right),$$

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

From now on, we shall use Equations 1.1 and 1.2 to refer to the BRDF formulation, for the sake of brevity.



(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 1.2: 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.

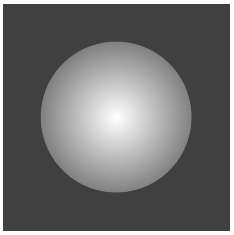
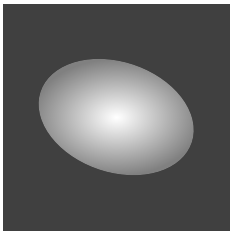
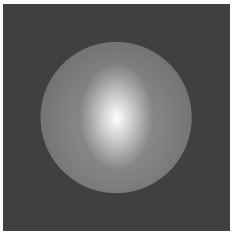
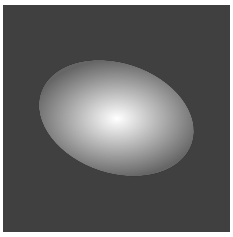
Surface	Straight	Rotated w.r.t. z- and y-axes $(30°, 15°)$
Lambertian		
Non-Lambertian		

Table 1.3: Different reflections on a flat circle-shaped object from different viewpoints. A visibly clear local difference in brightness and colour can be observed in between the viewpoints on a non-Lambertian surface, which leads to possible false or missing feature matches for feature matching algorithms relying on colour or brightness values.

Further generalizations of the BRDF model have been presented, such as 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; and a bi-directional surface scattering reflectance distribution function (BDSSRDF) with eight parameters, to quantify surface's internal scattering [37]. 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 [49]. For example, the Lambertian reflectance is a special case, where f_r is a constant function [53].

The main observation from the model is, however, that non-Lambertian surfaces' specular reflections — called *specularities* from now on — are dependent on a viewing angle, as the angle parameters change depending on the viewing direction as well. This phenomenon is further illustrated in Table 1.3. When a non-Lambertian surface is tilted with respect to y- and z-axes, while the incident light's direction stays constant, a visible local difference in brightness and colour can be observed. The exact nature of these specularities can vary, as the specularities can be local and sparse, like little pinpricks, or global and dense, which are blindingly bright areas. Regardless of the nature of these specularities, the non-Lambertian reflectance violates the following assumptions:

i) brightness constancy, where it is assumed that the brightness of pixels stays constant across images and

ii) colour constancy, where it is assumed that the colour of objects stays constant across images [58].

Due to these violations, non-Lambertian surfaces can, for example, lead to false image patch feature matches 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 [73]. Furthermore, as depth estimation algorithms rely on sufficient texturing, and specularities cause the regions to look uniformly textured despite the different ground truth, the depth and surface reconstruction can also be erroneously estimated for non-Lambertian surfaces [53]. Another notable subclass of problems arises in medical computer vision, where global and dense specularities occur in closed non-adjustable environments of human bodies during medical endoscopies [53]. 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., Mirko et al., and Saint-Pierre et al. respectively [38] [39] [50]. These far-reaching complications along with the computer vision's natural interest in optical exceptions has led to versatile research about non-Lambertian surfaces, compassing numerous fields of computer vision. Let us now delve into them.

2. Non-Lambertian surfaces

The major fields actively studying exclusively non-Lambertian surfaces are quite certainly related to *visual rendering*, such as photometric stereo and shape and texture from shading. The main reason for this are the common research questions posed by them, such as the accurate 3D reconstruction of objects under different, possibly unknown, lighting conditions and the extraction of surface normal maps from objects [53]. As such, non-Lambertian surfaces pose an immediate problem, as the specularities block the information that is meant to be extracted and thus there is an acute need for a method dealing with them. The method in turn varies depending on the exact research question and test setting of the study. Let us then investigate the setting and the research questions in detail. I shall consider mostly photometric stereo as it is closely coupled with shape and texture from shading in general.

2.1 Photometric stereo

In photometric stereo, there are multiple adjustable light sources present in the environment [58], and the research question is generally formulated as modelling the specularities' behaviour under different lighting conditions [53]. Likewise the name "photometric stereo", coined by Woodham in 1980 [67], alludes to the multiple light sources present in the environment taking up the role of image sensors in a traditional stereo system. Additionally — as opposed to the term "stereo camera" referring to the camera's characteristics — the usage of different sensors, such as intensity and image sensors, is not out of the question in the studies of photometric stereo [53].

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 1.1. Assuming this model in the form presented in Equation (1.1), 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

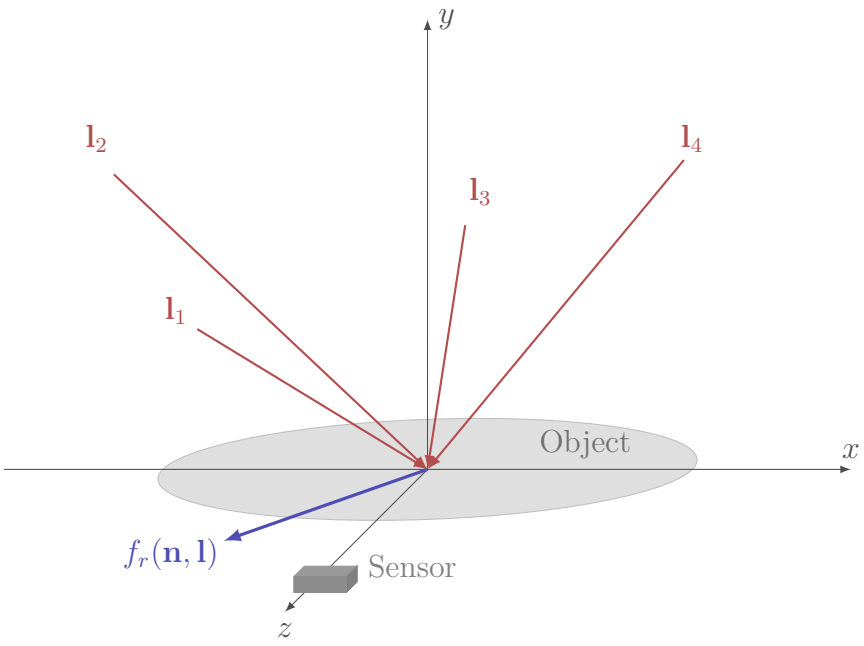


Figure 2.1: An illustrative diagram of a general test set-up in photometric stereo. Red arrows indicate incident lights and the blue arrow is the reflected light encoded by the BRDF. A dark gray box indicates the image or light sensor, and the light gray circle the object surface.

multiplication,

$$\mathbf{I} = \max\{f_r\left(\mathbf{n}, \mathbf{l}\right) \circ \left(\mathbf{N}^T\mathbf{L}\right), 0\}, \tag{2.1}$$

by using different assumptions and constraints on the $m \times k$-dimensional BRDF, $f_r(\mathbf{n}, \mathbf{l})$ [53]. Respectively in *uncalibrated photometric stereo*, the matrix $\mathbf{L}$ is unknown, and it needs to be estimated before solving the $\mathbf{N}$ [53]. Here it is good to take note that the zero, the second argument of the max function, represents the apparent shadow of the surface [53], which doesn't give up any information about the BRDF. The environment and test set-up is illustrated in Figure 2.1.

The logical next question is what kind of different assumptions and constraints might be of use to us. One is naturally the Lambertian reflectance model that was mentioned before, mathematically put as

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

where $\mathbf{D}$ is a diagonal matrix with each row being a constant, representing the constant diffuse radiance [53]. 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 $\mathbf{N}$, as proven by Woodham in 1980 [67]. 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 [68]. Additionally it does not apply outdoors due to sun's nigh planar trajectory causing the inverse of the $\mathbf{N}$ to disappear [67]. 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 [18]. Thus more recent efforts focus on a general unknown f_r that doesn't directly comply with the Lambertian reflectance model [53].

Another useful assumption would then be the local and sparse nature of specularities and shadows, which gives us a possibility to detect and discard them as outliers. These *outlier rejection methods* have been studied as early as 1982 by Coleman and Jain [11], and as such, there exists various criteria for detection and rejection. A few examples include studies by Verbiest and Van Gool from 2008 [61], and Wu and Tang from 2009 [68], which model inliers or outliers 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, such as shadows, noise and specularities, to form a sparse matrix $\mathbf{E}$, which is added to the Lambertian reflection matrix $\mathbf{D}$ [53]:

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

Consequently by minimizing the rank of $\mathbf{E}$ 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 $\mathbf{D}$ — we can achieve more robust rejection of specularities [53]. An example of this approach can be made from the study of Ikehata et al. from 2012 [27], where a hierarchical Bayesian approximation is used to estimate surface normals while modeling the $\mathbf{E}$ 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 [27].

The inherent weakness of any outlier rejection method lies in the implicit assumption of local and sparse outliers: when met with dense outliers, the algorithms' accuracy decreases [53]. For Ikehata et al., additional problems arise from non-Lambertian diffusive surfaces that don't fit in the statistical model [27]. 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, both Verbiest and Van Gool [61] and Wu and Tang [68] use a dense set of images, which translates to having over 100 images per reconstructed object. Henceforth, while being robust methods, the research in photometric stereo has veered towards *analytical BRDFs* accounting for outliers as well [53].

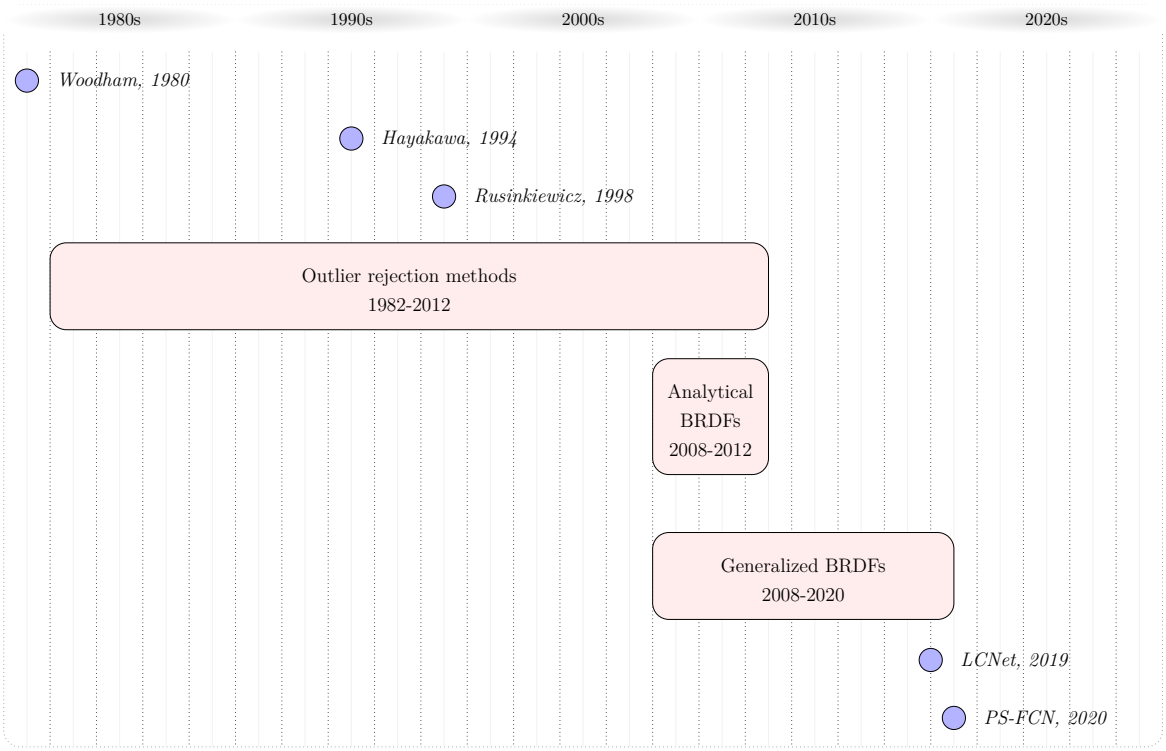


Figure 2.2: An approximate timeline of the milestones in photometric stereo, regarding non-Lambertian surfaces.

A couple analytical BRDFs include the Ward and Sparrow-Torrance models [53], which model the surfaces as a set of microfacets, microscopic surface areas acting as individual specular reflectors [46]. The distribution of microfacets' normals then differ from the surface normal depending on the surface's characteristics specified by the model [46]. For example, the Torrance-Sparrow model assumes microfacets to be perfectly specular and thus only the microfacets with their normal equal to the half-angle vector $\mathbf{h}$ can cause specularities in the viewing direction [46]. This model has been adapted, for instance, in a study of Georghiades regarding the uncalibrated photometric stereo in 2003 [21]. In general though, more studies have been dedicated to the Ward model, which assume an elliptical Gaussian distribution for the isotropic microfacet normals, thus having no preference over the reflectance direction [65]. The studies by Chung and Jia in 2008 [10], Goldman et al. in 2010 [22] and Ackermann et al. in 2012 [1] all use the Ward model or a variation of it. The exact approach to the outliers in the Ward model varies: Chung and Jia use the shadows to estimate the parameters of the BRDF [10], whereas Goldman et al. optimize the object shape and model parameters, and pixel-wise parameters and surface normals in alternating turns [22], and finally Ackermann et al. select the less shadowed pixels, which are most likely

to offer viable info about the BRDF [1].

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, the models can be non-linear, requiring careful and long optimization [53] and finally, there is no guarantee that an analytical model fits the observed BRDF well [49]. 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 leads to solving a non-linear equation, with the estimation of the surface normal at the same time [22].

Thus 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 [53]. These *generalized BRDFs* are further supported by the fact that materials often show structured BRDF values in real life, implying isotropy [53]. Isotropy then simplifies the mathematical formulation of the BRDF in a half-angle coordinate system, presented in Equation (1.2): the function has now only three parameters, as the ϕ_h is no longer necessary [63]. Monotonicity, in the other hand, implies that the intensity increases as the input increases in value, giving a unique inverse function for the BRDF [63]. 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 by Rusinkiewicz in 1998 [49]. The bi-polynomial model is then formulated in the following fashion [49]:

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

An example of a bi-polynomial model is the study of Shi et al. from 2014 [54], where it is further assumed that the aforementioned equation can be factored into two separate terms $g_1(\theta_h)$ and $g_2(\theta_d)$, enabling 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 was used in the study of Wang et al. from 2020, where the incident light is assumed to be collocated with the viewing point, allowing to decouple the surface normal from the BRDF [63]. 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 [53].

Another avenue implicitly used both in generalized and analytical BRDFs is *the component-wise structure of the reflection* [49]. 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

General equation for photometric stereo: $\mathbf{I} = \max\{f_r(\mathbf{n},\mathbf{l}) \circ \left(\mathbf{N}^T\mathbf{L}\right), 0\}$			
Proposed solution	Study	Assumptions	Weaknesses
$f_r(\mathbf{n},\mathbf{l}) \approx \mathbf{D}$, linear least square	[67]	Lambertian reflection, a constant diagonal $\mathbf{D}$	non-Lambertian reflections
$f_r(\mathbf{n},\mathbf{l}) \approx \mathbf{D} + \mathbf{E}$, outlier rejection	[61], [68]	Local and sparse specularities	Global and dense specularities, a large number of images
Analytical BRDFs, eg. Ward and Torrance-Sparrow models	[10], [21], [22]	Microfacets and their surface normal distribution w.r.t. specularities	Computational complexity, material specificity
Generalized BRDFs	[18], [63]	Component-wise reflection or generic properties of BRDF	Reflections or components, which violate the assumptions, eg. anisotropic reflection
Bi-polynomial approximations $f_r(\mathbf{v},\mathbf{l}) \approx g(\theta_h,\theta_d)$	[54]	Monotonic and isotropic BRDF	Reflections or components, which violate the assumptions, eg. anisotropic reflection
Specialized data collection methods	[34], [55], [76]	-	Unpractical set-up, lack of robustness and generality
Depth priors	[23], [75]	-	Definition and collection of priors, computational complexity
Masking methods	[38], [39], [50]	Uniform colour or limited colour variation	Lack of robustness and generality, soft specularities
Neural networks	[6], [30], [59]	-	Lack of a well-defined BRDF, explainability and various special cases

Table 2.1: A table describing briefly the proposed solutions in calibrated photometric stereo.

models [18]. An example of this approach can be seen in the study of Earp et al. from 2007, where a pseudo-specular and diffuse Lambertian component is used [18]. 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.

The most recent venture in photometric stereo involves using *neural networks* that learn the reflectance mapping directly [6]. While this approach tends to have the advantage of dealing with majority of reflectance at ease, a few limits exist: Chen et al. experience significant error with noisy light intensities [6], Taniai et al. have a long runtime and decreased performance in complex reflections [59] and Kaya et al. have problems with concave shapes [30].

As we have now covered most of calibrated photometric stereo, let us now cover uncalibrated photometric stereo as well. As noted earlier, uncalibrated photometric stereo has to estimate the lighting matrix $\mathbf{L}$ in Equation (2.1) along with the surface normal matrix $\mathbf{N}$ [53]. 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 [24]. As the normals are scaled, we can solve the ambiguity caused by scaling, denoted as $\mathbf{A}$, by the singular value decomposition [24] or matrix factorization [15]. Thus the whole problem can be stated as the following equation,

$$\mathbf{I} = \max\{\mathbf{D} \circ \left(\mathbf{N}^T\mathbf{L}\right), 0\} = \mathbf{S}^T\mathbf{L} = \hat{\mathbf{S}}^T\mathbf{A}^T\mathbf{A}^{-1}\hat{\mathbf{L}} \tag{2.2}$$

where $\hat{\mathbf{S}}$ and $\hat{\mathbf{L}}$ are the unscaled pseudo-normal and pseudo-lighting matrices respectively [24]. 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 [53]. For instance, the rotation ambiguity can be solved with six different surface points with constant albedo or intensity [53], but more common constraints are the integrability of the surface or the observation of shadow boundary, which reduces the problem to *the Generalized Bas-Relief ambiguity*, stated as the following equation [24]:

$$\mathbf{I} = \max\{\mathbf{D} \circ \left(\mathbf{N}^T\mathbf{L}\right), 0\} = \mathbf{S}^T\mathbf{L} = \hat{\mathbf{S}}^T\mathbf{A}^T\mathbf{A}^{-1}\hat{\mathbf{L}} = \hat{\mathbf{S}}^T\mathbf{G}^T\mathbf{G}^{-1}\hat{\mathbf{L}}. \tag{2.3}$$

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 [53]. Each approach comes with their own limitations and advantages. For contrast, Shi et al. use chromatic clustering to detect points, which have equal albedo [52], Papadhimitri and Favaro locate the points where $\mathbf{n} = \mathbf{l}$ [44], and Alldrin et al. minimize the entropy after assuming a limited amount of dominant colours in the image [2]. The first study is unsuitable to grayscale images [52], the

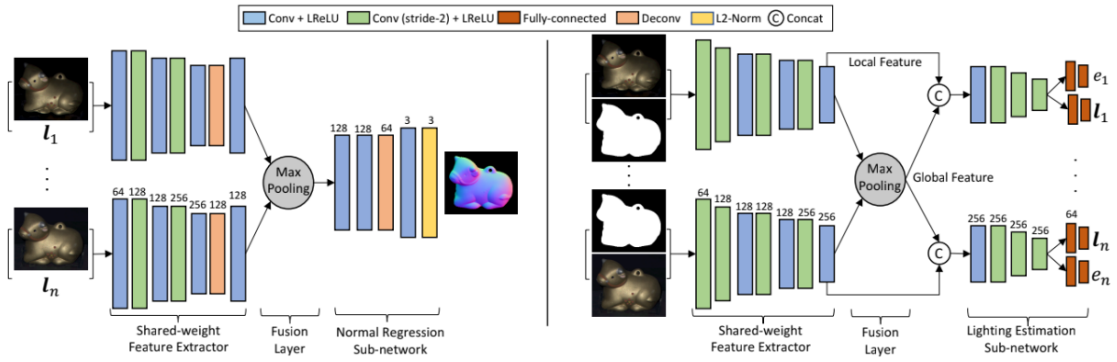


Figure 2.3: A diagram of the architecture of PS-FCN (left) and LCNet (right). Cited from Chen et al. [6].

second is limited solely to the diffuse component [44] and the third requires intricate pre-processing steps [2].

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 [7]. A notable example is the LCNet from 2019 by Chen et al. [7] that uses convolutional layers and max pooling to detect global features from local features. The network has been further enhanced in the study from 2020 by Chen et al. [6] and it ranks among state-of-the-art systems in photometric stereo. The architectures of both networks, Photometric Stereo Fully Convoluted Network (PS-FCN) and Light Calibration Network (LCNet), are illustrated in Figure 2.3.

Another study using neural networks has been conducted by Kaya et al. in 2021 [30]. Both still have performance issues with ambiguous special cases: Chen et al. with piece-wise planar and planar surfaces with uniform albedo [6] and Kaya et al. with concave shapes [30]. Another avenues outside the Equation (2.2) 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 [53]. Examples include the study by Sato et al. from 2007 [51] and Lu et al. from 2015 [36].

Finally, aside from solving the Equation (2.1) by any of the aforementioned approaches, there are numerous methods of solving calibrated or uncalibrated photometric stereo by changing the data collection method or the input [53]. For example, a study by Zhou et al. uses multi-spectral light field, which gives them additional constraints of multiple viewpoints and point lights and thus more accurate measurements of the surface normals' orientation [76]. Object motion works in a similar albeit stricter manner [53], and it has been utilized in the studies of Simakov et al. in 2003 [55] and Lim

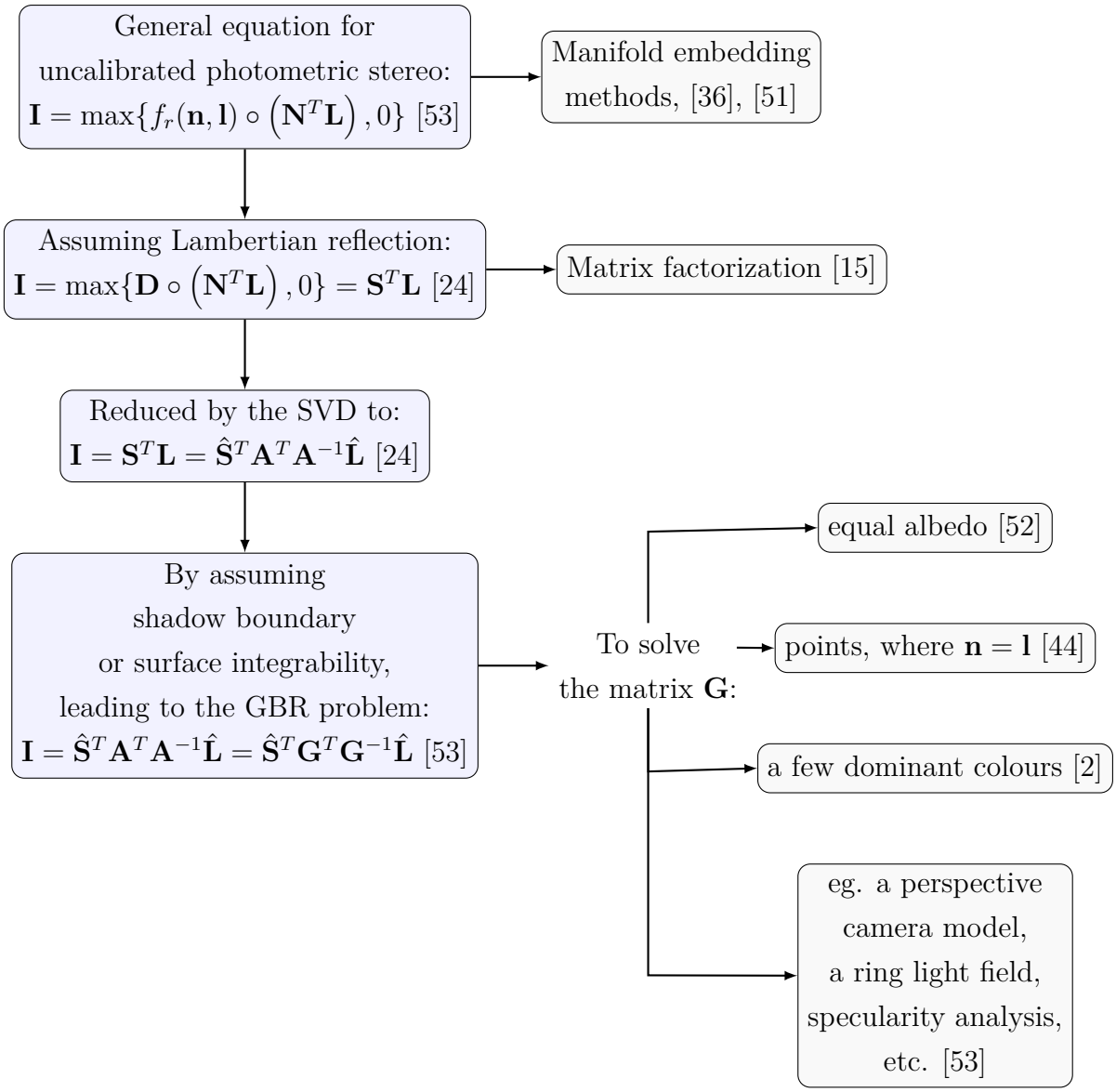


Figure 2.4: An approximate diagram describing briefly the proposed solutions in uncalibrated photometric stereo. Blue boxes mark the steps required for the solutions on the right, drawn in light gray boxes.

et al. in 2005 [34]. An alternative solution can be found from the colour channels, which reveal specularities when studied individually [53]. This approach has also been utilized to create sophisticated *masking methods* for medical endoscopies in the studies of Saint-Pierre et al., Mirko et al., and Meslouhi et al. respectively between 2007 and 2011 [38] [39] [50]. 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 [53], like has been done in the studies by Zhang et al. in 2012 [75] and Hague et al. in 2014 [23]. 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 [53]. 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 of photometric stereo. The approximate timeline for the milestones in photometric stereo is presented in Figure 2.2, and Table 2.1 and Figure 2.4 recount the presented solutions to calibrated and uncalibrated photometric stereo in a general level respectively. As we have now gone over photometric stereo in an extensive manner, it is time to ask what might the other fields of computer vision offer regarding the non-Lambertian surfaces. One worth taking a good look at is the field of navigation, where the non-Lambertian surfaces play a significant role in a central concept of featurelessness. Let us go there next.

2.2 Navigation

Navigation has been around since 1980s in computer vision research [58], entertaining the ideas of an autonomously moving robots and vehicles. As Szeliski points out, the research has progressed rapidly in various areas, producing a new golden standard for today: *Simultaneous Localization And Mapping*, shortly put SLAM. What makes SLAM different from its predecessors is its unique approach to the key questions contained in its name, localization ("where we are") and mapping ("what is around us") solved at the same time. Consequently, the availability of different sensors has been a major driving force in the emergence of new SLAM methods [5]. Let's present now the most basic mathematical formulation of the SLAM. Given

- $t+1$ discrete time steps, often realized as frames of the camera or the frame rate of the video,
- a t-dimensional vector $\mathbf{o} = (o_1, o_2, \ldots, o_t)$ of previous observations and the current observation o_{t+1}, usually in the form of image or inertial data or both,
- a t-dimensional vector $\mathbf{c} = (c_1, c_2, \ldots, c_t)$ of control inputs associated with each time step, such as the vehicle state or camera calibration parameters,
- a t-dimensional vector of previous locations, denoted as $\mathbf{x} = (x_1, x_2, \ldots, x_t)$ and
- a t-dimensional vector of previous environment maps, denoted as $\mathbf{m} = (m_1, m_2, \ldots, m_t)$ and practically translated as anything of interest in the en-

vironment, such as landmarks, objects, vehicle's relative location or background images,

we need to predict the location and environment map at the current time step, denoted x_{t+1} and m_{t+1} respectively [17], as demonstrated in Figure 2.5. The SLAM in early days was a probabilistic problem, so assuming this formulation, we have the following equation as our objective,

$$P(x_{t+1}, m_{t+1}|\mathbf{c}, \mathbf{o}, \mathbf{x}, \mathbf{m})$$

and we need the predict the joint posterior probability distribution at each time step [17]. Assuming the Markov property for locations — that is, the next location x_{t+1} is only dependent on the previous location x_t and control input c_t — and the conditional independence of observations given the environment and current control inputs, applying Bayes' theorem makes the problem recursive for observations and environment maps [17]. In this form, the problem can then be formulated followingly:

$$P(x_{t+1}, \mathbf{m}|\mathbf{o}, \mathbf{c}, x_0) = \int P(x_{t+1}|x_t, c_{t+1})\, P(x_t, \mathbf{m}|\mathbf{o}, \mathbf{c}, x_0)\, dx_t \tag{2.4}$$

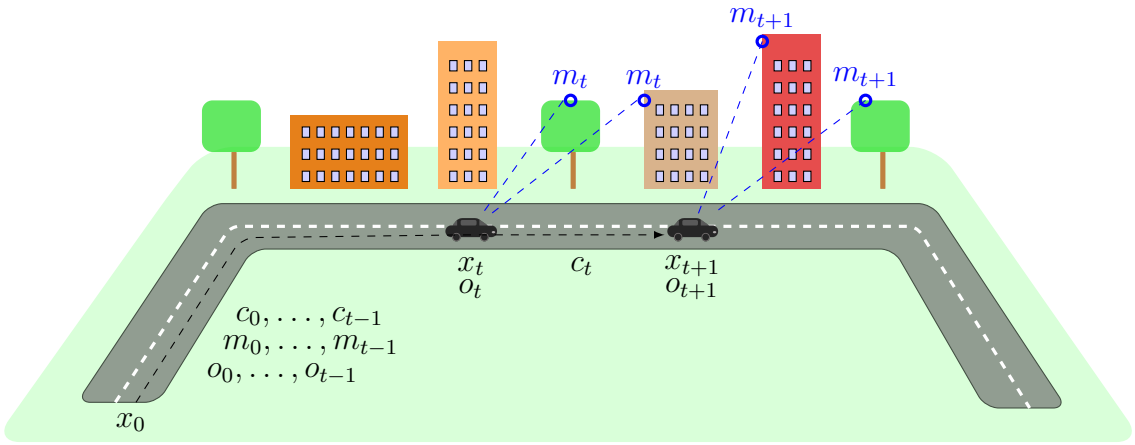
$$P(x_{t+1}, \mathbf{m}, m_{t+1}|\mathbf{o}, o_{t+1}, \mathbf{c}, x_0) = \frac{P(o_{t+1}|x_{t+1}, \mathbf{m})P(x_{t+1}, \mathbf{m}|\mathbf{o}, \mathbf{c}, x_0)}{P(m_{t+1}|\mathbf{o}, \mathbf{c})}. \tag{2.5}$$

These equations can be further refined depending on, for example, whether we assume the location x_t is known, leading to a conditional probability density function, or the environment maps' locations are known, when the objective changes to predicting the vehicle's relative location with respect to environment maps [17]. However, the aforementioned Equations (2.4) and (2.5) are sufficient for the purpose of our literature survey, so we shall not make further assumptions and instead move on to investigate how the SLAM solves this problem in general and how non-Lambertian surfaces play a part in it.

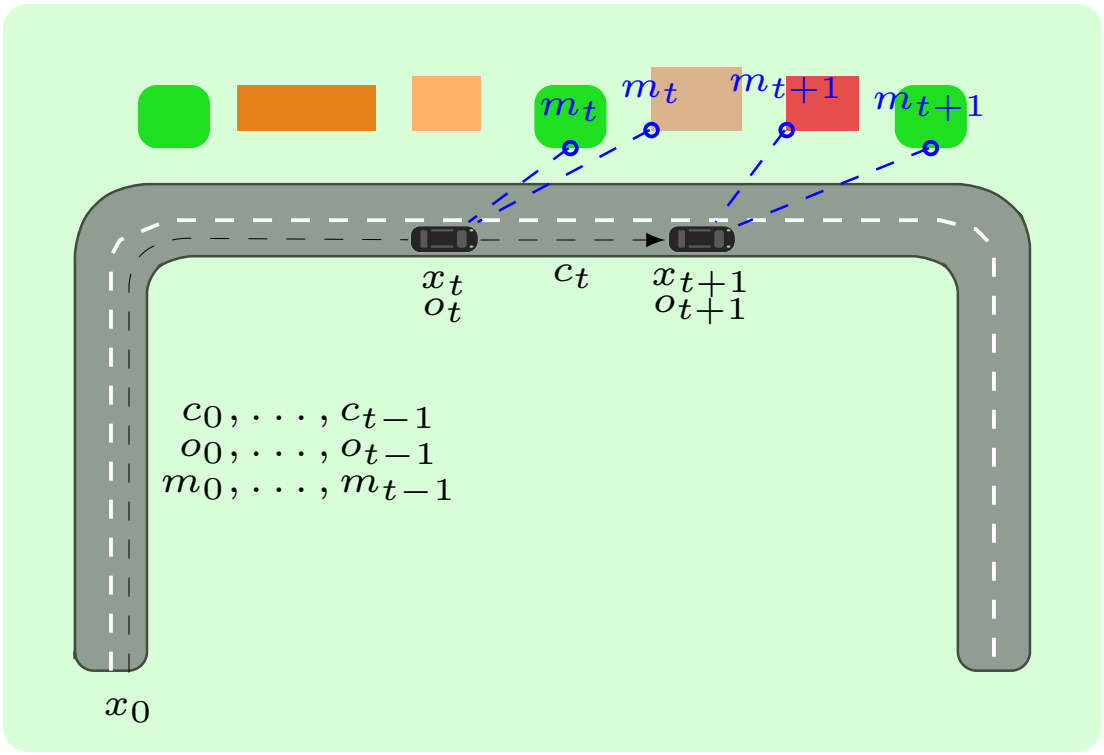
In a probabilistic form, the solution for SLAM is to find the transition models for observations and environment maps, commonly referred to as *localization* and *mapping* respectively as per the name SLAM [17]:

$$P(o_{t+1}|x_t, \mathbf{m}), \quad P(x_{t+1}|\mathbf{x}, c_t). \tag{2.6}$$

General approaches how to construct these models can be categorized into three different groups based on their sensors and input data: *RGB-D SLAM*, where both monocular RGB cameras and depth sensors are used; the *visual-inertial SLAM* (VI-SLAM), where the inertial measurement units (IMUs) and images from stereo or monocular cameras are used; and the *visual SLAM*, where only the stereo or monocular images are used [5]. Another categorization of SLAM is a division into *feature-based* (indirect) and *direct methods*. The feature-based methods use only features consisting of image



(a) A 3D view.



(b) A 2D view.

Figure 2.5: 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 environment maps along the way. The control variables (driving cues) at a time step i are denoted by c_i, the input data is denoted by o_i, the previous environment maps by m_i and the car's location by x_i.

keypoints and their respective descriptors to extract camera poses, whereas the direct methods use the sensor data without pre-processing, minimizing the photometric error in the environment maps [5]. It is good to note here that the categorizations of environment mappings' density and input data are not mutually exclusive here, and thus any combination of them is possible, resulting in eight options. This difference is illustrated further in Figure 2.6.

The key observation in approaches besides their input is the density of the environment maps [5]. Depending on the approach, desired computational complexity and research question, *dense, semi-dense* or *sparse environment maps* nay be sought, as the density poses a significant computational constraint [5]. In the other hand, feature-based methods experience poor performance in featureless environments, as it generates a sparse map only [5]. The visual difference between different maps is illustrated in Figure 2.6. Here we can observe that non-Lambertian surfaces are thus ideal to cause problems for feature-based methods, but the issue can not straightforwardly be excluded for the direct methods either: the unprocessed data can consider also pixels' intensities, where the brightness constraint is of utmost importance [5].

Another key issue of non-Lambertian surfaces then regards the tracking process, as the environment maps affect them as well as a joint optimization problem. I shall use a concept called *optical flow* to highlight the root of the issue, but it is quintessential to note that while optical flow is a prevalent approach in tracking, it is not the only possible approach. Thus non-Lambertian surfaces may pose different issues for alternative approaches, but for the purpose of this survey and understanding the issue posed, it is sufficient to regard this one example.

Optical flow is used to model the relative motion between the sensor and object using constant brightness values across the frames, formulated as

$$I(x, y, t) = I(x + \Delta x, y + \Delta y, t + \Delta t) \tag{2.7}$$

where a pixel at spatiotemporal location (x, y, t) with intensity $I(x, y, t)$ has moved by $\Delta x, \Delta y$ and Δt between the two frames [58]. As non-Lambertian surfaces violate the brightness constancy assumption, the above Equation (2.7) does not hold and the optical flow is disturbed. This issue is then manifested as a phenomenon called *drifting*, the gradually occurring deviation between the actual and predicted location of the vehicle [17]. Generally formulated outside the framework of optical flow, non-Lambertian surfaces exhibit temporary featurelessness at a time step t in an observation o_t where we do not expect it to be, whereas in other time steps' observations specularities are not visible, thus affecting the system's localization via unpredictable and ambiguous input data.

Furthermore, as these locations are predicted constantly, the drift in one time step

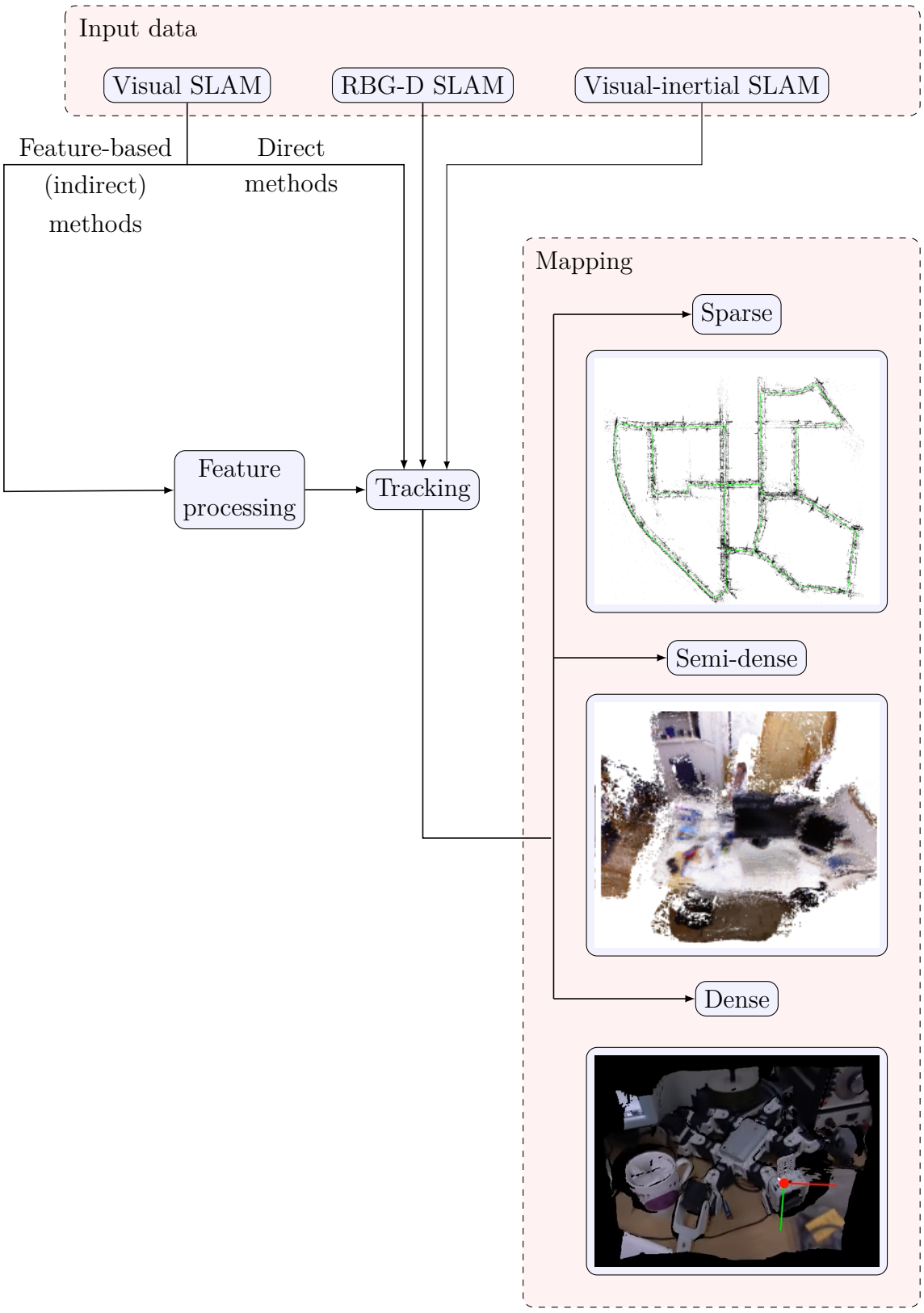


Figure 2.6: A diagram about the differences between SLAM systems. The sparse mapping has been cited from ORB-SLAM2 [41], the semi-dense from CNN-SLAM [60] and the dense from DTAM's presentation video [47].

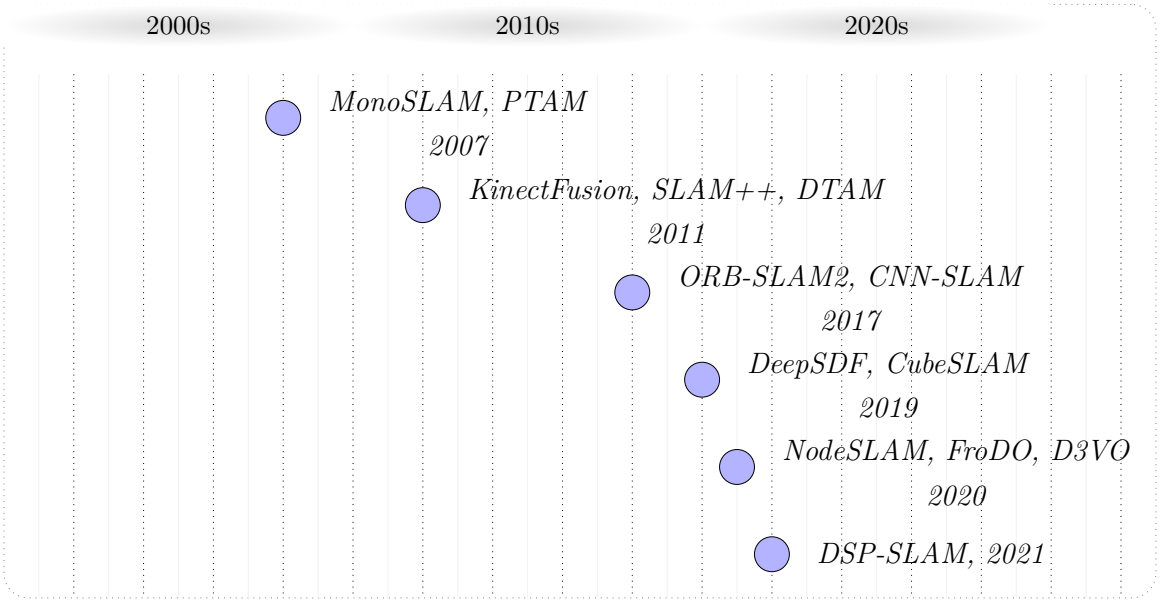


Figure 2.7: An approximate timeline of the milestones in navigation, regarding visual monocular SLAM and semantic SLAM.

will be implicitly inherited to the whole history of locations, commonly termed as the "trajectory" [5]. Ergo the drifting is an issue in terms of navigation system's overall accuracy, but it presents a substantial problem also for *loop closure*, a key process in SLAM [4]. To have an upper bound in the uncertainty regarding the transition models, SLAM systems correct their mapping and location prediction when entering a previously visited area [5]. The noteworthy point is that in the case of enough drifting, the detection of the familiar area might be skipped and thus the uncertainty and drift remains unbounded. While it can be noted that the loop closure is only present in SLAM systems out of all systems intended for navigation — indeed, the whole field of *visual odometry* is focused on localization predictions without the loop closure [58] — the drifting is ideally kept to minimum in every one for better performance and thus non-Lambertian surfaces are a notable and diverse issue in navigation as well.

Consequently, a sizeable effort has been dedicated to solving this problem [5], so let us recount the presented solutions. For the sake of completeness, I shall also present other milestones relevant for the empirical studies, presented later in Chapter 3, primarily those of visual monocular SLAM as the inertial input can be loosely coupled with the pose estimation of visual SLAM systems, thus making the boundary between the VI-SLAM and visual SLAM flexible [5].

The very first SLAM study to propose an algorithm for monocular cameras was the study of Davison et al. in 2007 named MonoSLAM [14]. Besides the unprecedented equipment, the novelty of this study lies in the approach used to compute the movement

between the time steps, namely the combination of "smooth camera movement" and the initialization procedure. In other words, it is assumed that the camera is moving at constant velocity in between the frames and the first environment map is initialized on start-up, by introducing a known object to the first image. Then by adopting the probabilistic formulation of SLAM, the solution presented by Davison et al. was an extended Kalman filter, a probabilistic non-linear model that uses measurements o_t to correct the existing predictions x_t, as illustrated in Figure 2.8. In this solution, the transition models of locations and observations, denoted by f and h respectively, are non-linear yet differentiable, formulated in a following fashion:

$$x_t = f(x_{t-1}, c_t) + w_t$$
$$o_t = h(x_t) + v_t,$$

where w_t is the process noise at the time step t and v_t is the measurement noise at the time step t. The drawbacks of MonoSLAM are the high memory costs of a feature-based approach, and the lack of global optimization to reduce this load. With a memory complexity of $O(n^2)$, the size of the environment determines the usability of the algorithm: Davison et al. themselves only used a sparse environment map of 100 features in a single room, where the robot's trajectory was a circle with a 1.5 diameter as to preserve the frame rate at 30 per second.

At around the same time, monocular feature-based PTAM was also introduced by Klein and Murray, which had a new approach to global optimization: the decoupling of the feature tracking and environment mapping to different threads [31]. This, along with the idea of expanding the initial map gradually with new frames of interest, allowed for thousands of features in the environment maps. The resulting environment maps, however, were still low-quality point clouds, thus at best when used with landmarks. Additionally, PTAM was not equipped to handle efficiently loop closure or significant occlusion, ie. blocking of vision caused by invisible object. The architecture of PTAM

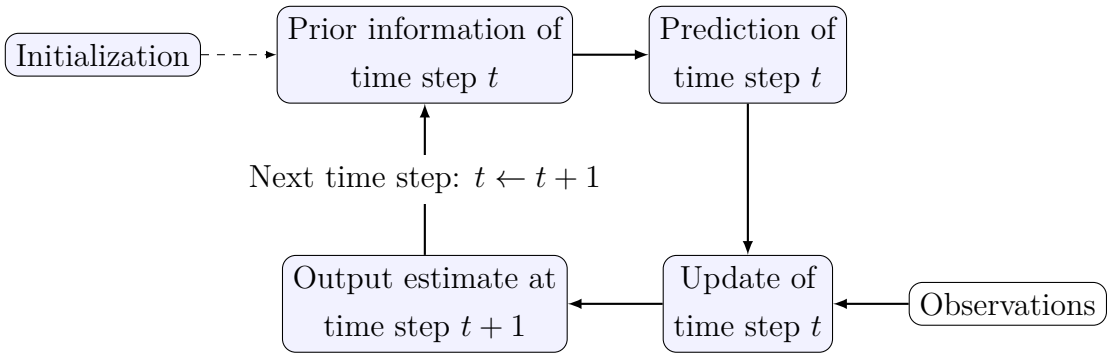


Figure 2.8: A diagram of an extended Kalman filter pipeline used in MonoSLAM. Blue boxes marks the steps required, while the gray box denotes the data.

is illustrated in Figure 2.9.

The next world-changing study was KinectFusion, made by Newcombe et al. in 2011 [42]. This study, along with monocular cameras, sported a new type of low-cost depth sensor called "Kinect", which utilized infrared ranging techniques, making this a RBG-D study. These kind of depth sensors are nowadays referred to as light detection and ranging sensors, shortened as LiDAR sensors, [58], and are still widely used in many studies despite the different types of algorithms needed to process the data [5]. This led to advent of accurate ground truth depth information to be used alongside the following formula,

$$d_x x' = Kx,$$

which projects the monocular image point x' from a homogeneous coordinate system into a real-world point x, given the depth of the point x, denoted by d_x, and the camera intrinsic matrix, denoted by K [69]. KinectFusion also enabled the field of semantic SLAM, now already lurking around the corner.

Often heralded as the first semantic SLAM [76], SLAM++ was published in 2011 by Moreno et al [28]. 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 to build the object database required for the reconstruction. This 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, as can be seen from the studies of Chhaya et al. [9], Yang et al. [71], Häne et al. [26] and Dame et al. [12]. 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.

The next milestone is DTAM, Dense Tracking And Mapping, from 2011 by Newcombe et al., which was the first fully direct SLAM method ever, and is the first dense mapping method in our survey [5]. The mapping is based on a specialized loss function, which the authors call "the data cost volume" and is mathematically defined as the average photometric error over varying pixel-wise depths with respect to the current frame, calculated from the large collection of overlapping images [43]. The optimal solution minimizing the error is used as a regularizer for the environment mapping, along with a penalty function constraining the depth map to be spatially smooth. The tracking is in turn done by comparing the camera projection of the 3D environment mapping to the current frame, producing the estimated motion parameters. While the mapping and tracking are detailed and accurate, and robust to rapid movement, the

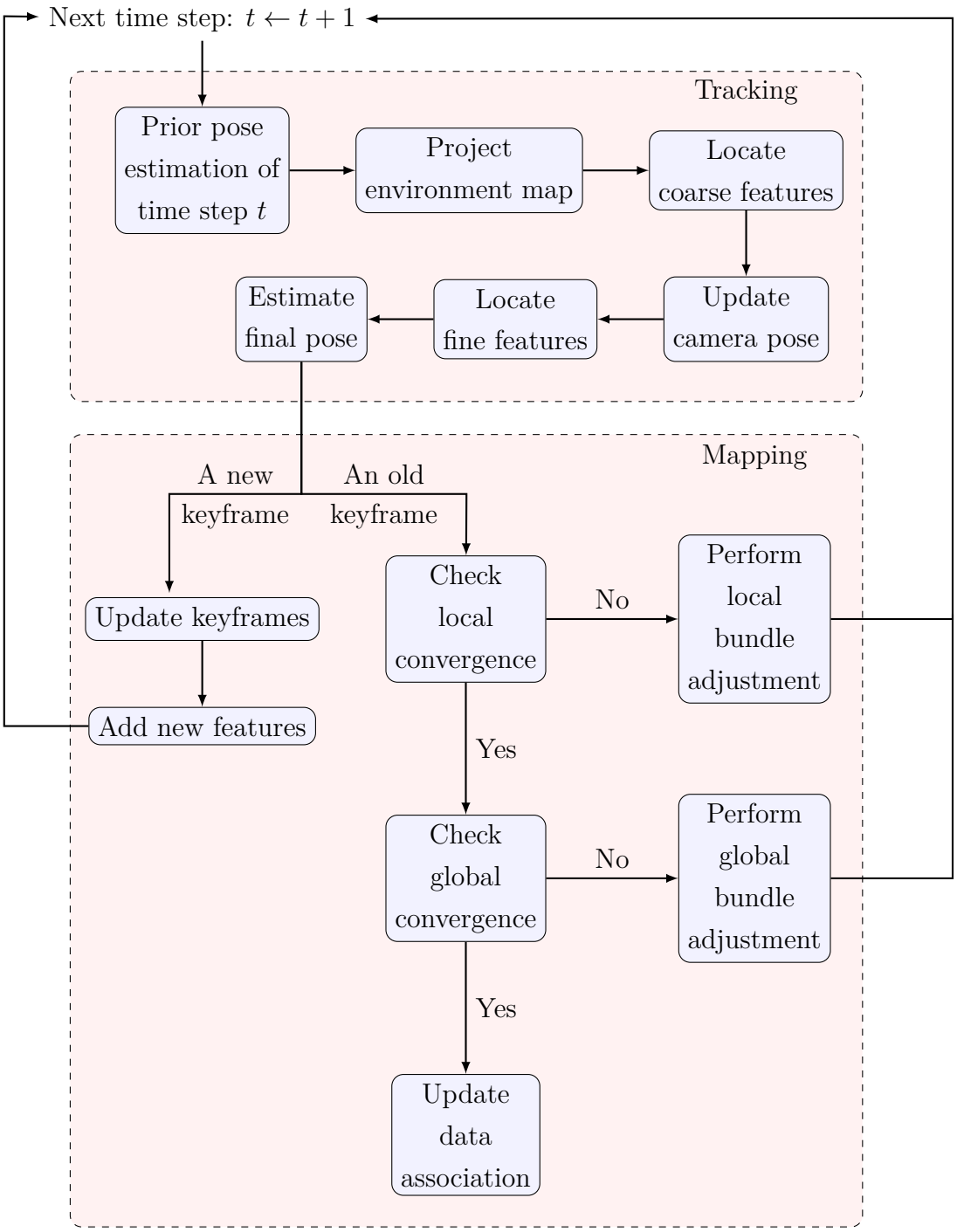


Figure 2.9: A diagram about PTAM architecture. Blue boxes marks the steps required, while red boxes indicate the threads involved.

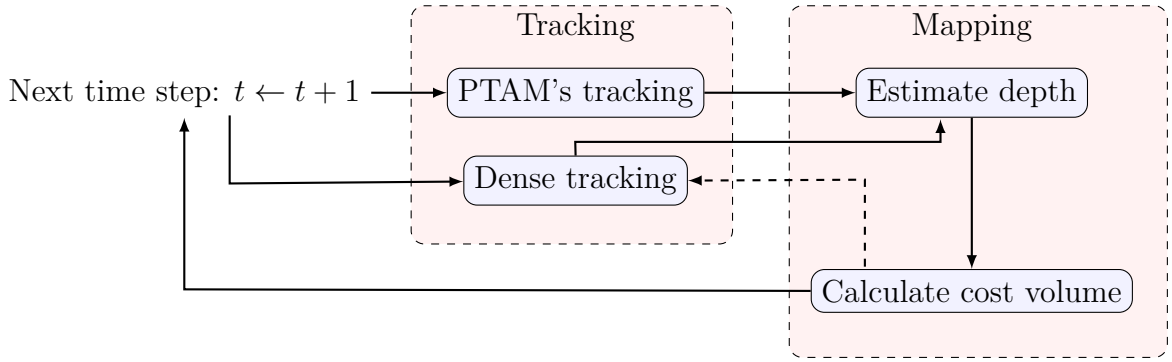


Figure 2.10: A diagram about DTAM architecture. Blue boxes marks the steps required, while red boxes indicate the threads involved.

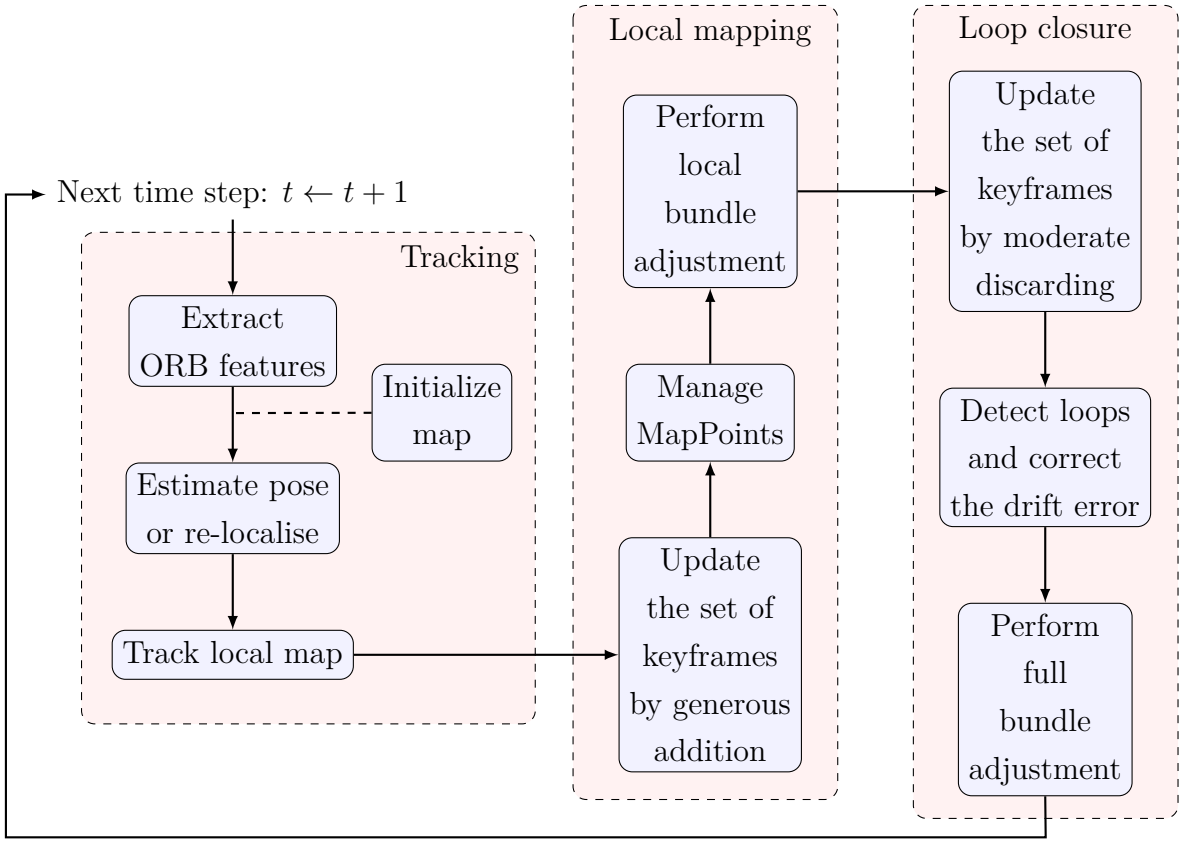


Figure 2.11: A diagram about ORB-SLAM2 architecture. Blue boxes marks the steps required, while red boxes indicate the threads involved.

advantages are paid for by the significant computational complexity and the lack of global optimization and loop closure techniques. The architecture of DTAM is illustrated in Figure 2.10.

The next SLAM method of interest is ORB-SLAM2 by Mur-Artal and Tardós

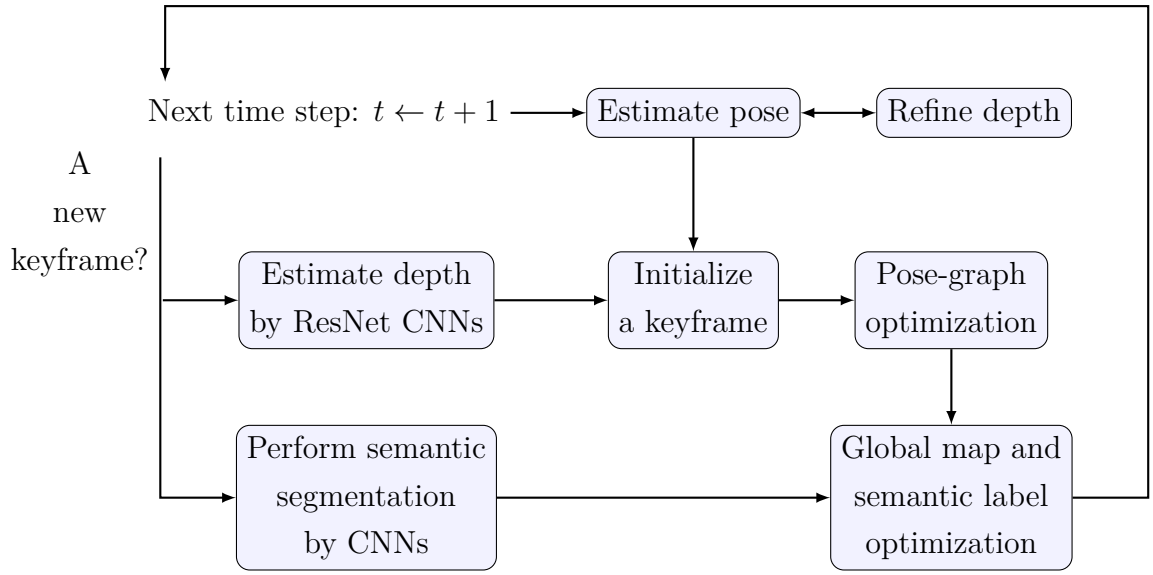


Figure 2.12: A diagram about CNN-SLAM architecture. Blue boxes marks the steps required. As the tracking and mapping threads are interleaved within each other, they are not marked in the architecture.

from 2017 [41]. It is even today considered as the state-of-the-art of feature-based SLAM systems due to its enhanced optimization and loop closure techniques, the wide range of applicable input from monocular and stereo images to RBG-D data [5], and a modern feature matching algorithm "Oriented FAST and rotated BRIEF" (ORB) by Rublee et al., which is built on the preceding algorithms of FAST and BRIEF, but with accelerated speed and better rotation invariance [48]. Utilizing its predecessor ORB-SLAM's ideas from 2015 [40], 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 [41]. The drawbacks of this widely used lightweight open-source solution stem from the weak robustness for motion blur [41] and featureless regions, which accumulate drift considerably in monocular input [25]. The architecture of ORB-SLAM2 is illustrated in Figure 2.11.

The next relevant milestone in our survey is the emergence of neural networks in SLAM systems around 2017 [5]. One of the first precursors was the CNN-SLAM, which utilized convolutional neural networks 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 [60]. The architecture is illustrated in Figure 2.12. Convolutional neural networks, shortly CNNs, continue to be widely used in semantic segmentation today [35], and indeed another type of neural networks is used in our next milestone study from 2019, DeepSDF

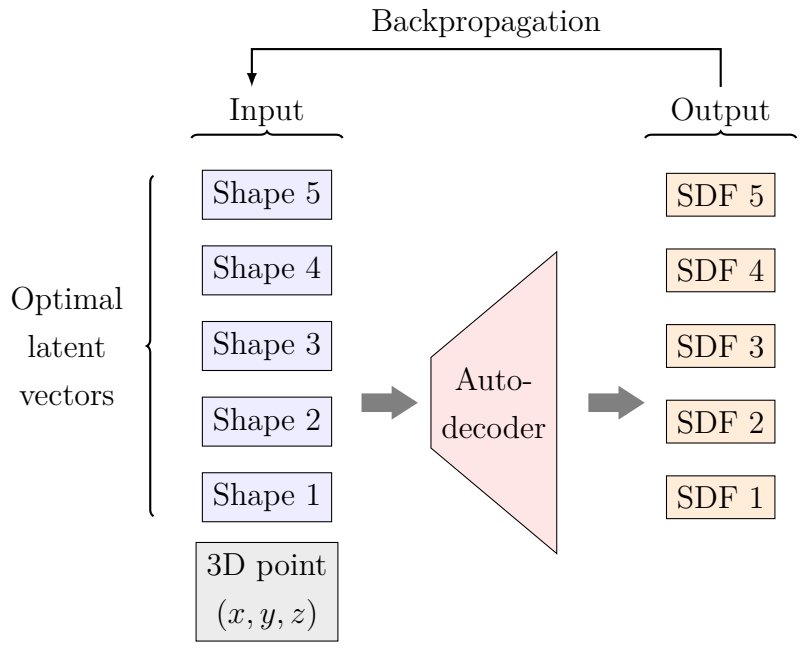


Figure 2.13: A diagram about the DeepSDF's architecture. Randomly initialized latent shape vectors (coloured blue) are input for the auto-decoder (coloured red), which optimizes them until convergence via backpropagation and outputs their SDF values (coloured orange) with respect to the other input, the 3D query point (coloured gray).

by Park et al. [45], which uses feed-forward networks in a probabilistic auto-decoder architecture to learn continuous signed distance functions (SDFs), as illustrated in Figure 2.13. A signed distance function f_d is defined as a continuous function of a point coordinate x, which outputs a real number s presenting the distance to a surface of interest in a following fashion:

$$f_d(x) = \begin{cases} s > 0, \text{ if } x \text{ is inside the surface} \\ s = 0, \text{ if } x \text{ is on the (decision) boundary of the surface} \\ s < 0, \text{ if } x \text{ is outside the surface.} \end{cases}$$

With the decision boundary of f_d the surface of interest can be constructed via ray-tracing or the marching cubes algorithm [45]. 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 following loss function,

$$\mathcal{L}(f_\theta(x), s) = |f_c(f_d(x), \delta) - f_c(s, \delta)|\,,$$

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

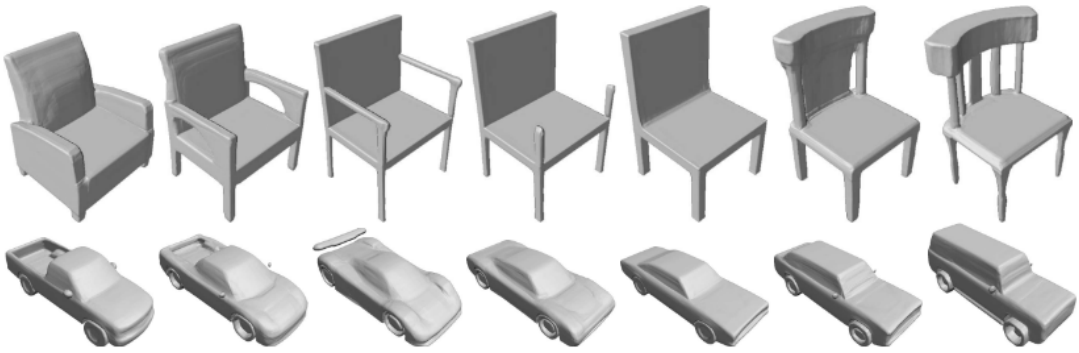


Figure 2.14: Raycast renderings of DeepSDF's converged latent shape vectors. Cited from Park et al [45].

metric SDF; and f_c is the real-valued "clamp function", defined followingly:

$$f_c(x, \delta) \coloneqq \min(\delta, \max(-\delta, x)).$$

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. Some of the optimized shape codes are visualized in Figure 2.14. 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.

A semantic SLAM we should take note of now is the CubeSLAM by Yang and Scherer, which unites the fields of monocular 3D object detection and SLAM systems once more [71]. 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. Continuing with semantic SLAMs, we have also NodeSLAM by Sucar et al. [57] and From Detections to 3D Objects (FroDO) by Rünz et al. from 2020. FroDO uses the DeepSDF and encoder architecture to further refine the monocular object detection via shape priors and other estimation steps [33], 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 [57].

The most remarkable study of 2020 is, however, D3VO by Yang et al. [70], which is a monocular visual odometry system — that is, a short-range monocular navigation system without loop closure. D3VO brings together the lessons of photometric

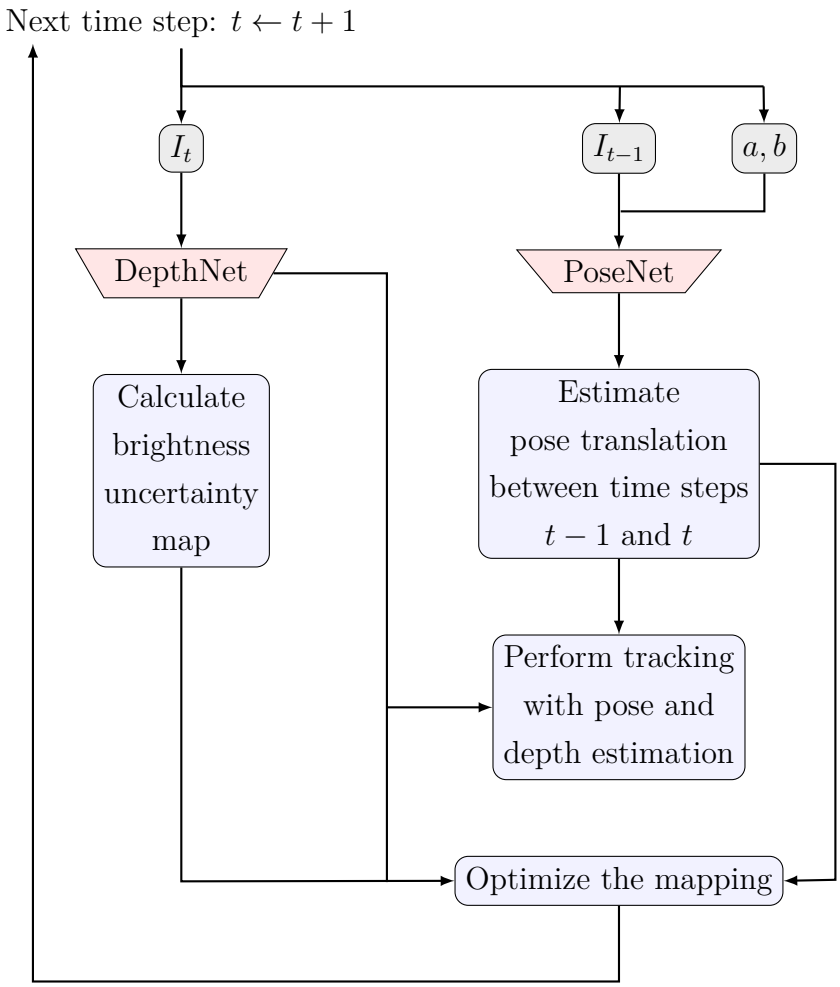


Figure 2.15: A diagram about D3VO architecture. Blue boxes marks the steps required, red trapezoids indicate the network structures in the system and the gray box denotes the input.

stereo and SLAM united under brightness affine transformation and deep learning [70]. 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, [70],

$$I_{t'} = aI_t + b, a > 0, b > 0$$

2. a photometric error with comparison functions l, c and s for luminance, contrast and structure [64],

$$r(I_t, I_{t'}) = \frac{\lambda}{2}\left(1 - \text{SSIM}(I_t, I_{t'})\right) + (1-\lambda)\left|\left|I_t - I_{t'}\right|\right|_1,$$
$$\text{SSIM}(I_t, I_{t'}) = [l(I_t, I_{t'})]^\alpha \cdot [c(I_t, I_{t'})]^\beta \cdot [s(I_t, I_{t'})]^\gamma,$$
$$\alpha > 0, \beta > 0, \gamma > 0, 0 < \lambda < 1,$$

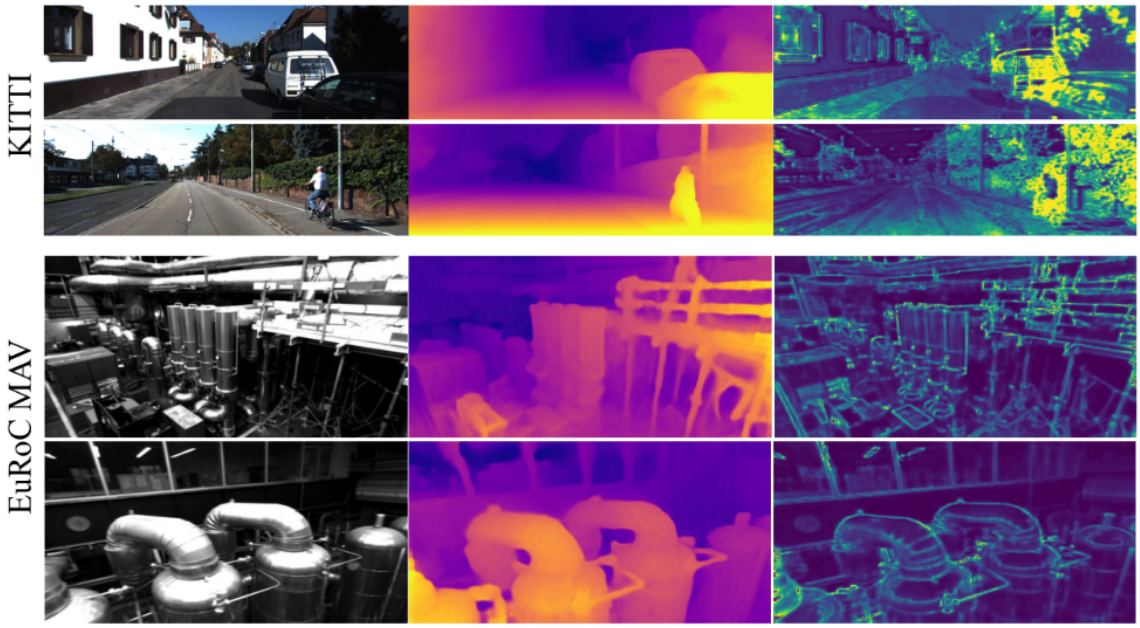


Figure 2.16: The original data (left), predicted depth image (center) and predicted uncertainty by D3VO on KITTI and EuRoC MAV datasets. Cited from Yang et al. [70].

and thirdly

3. an uncertainty map of true pixel intensity y with Laplacian noise [70],

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

and embedding all three of these equations into a self-supervised re-projection error [70], D3VO can then adjust the weighting of the residual for areas with high uncertainty, as demonstrated in Figure 2.16. This in turn achieves greater robustness against featurelessness [70]. The architecture of D3VO is illustrated in Figure 2.15.

Now we can move on to the final study in our survey. It is a state-of-the-art semantic SLAM by Wang et al. from 2021, called DSP-SLAM [62]. 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 low frame rate of 10hz, presented in Figure 2.18.

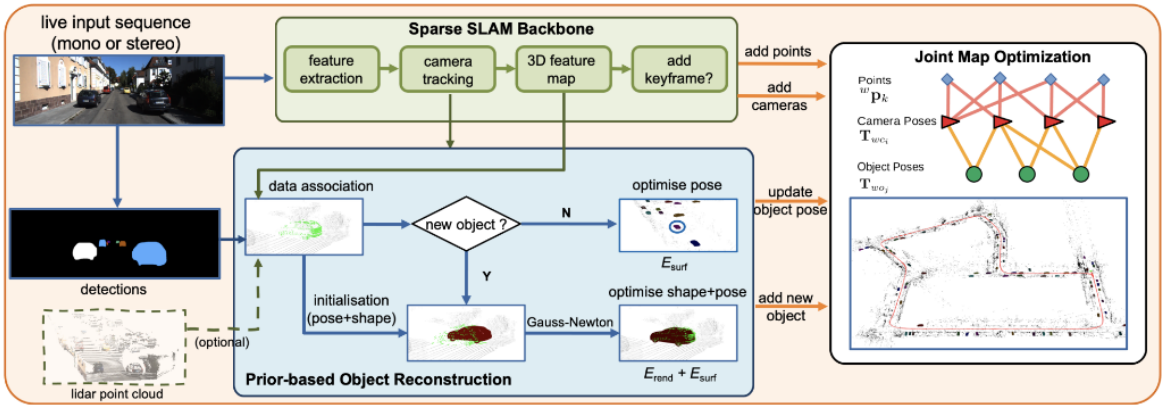


Figure 2.17: A diagram about the DSP-SLAM architecture. Cited from Wang et al. [62].

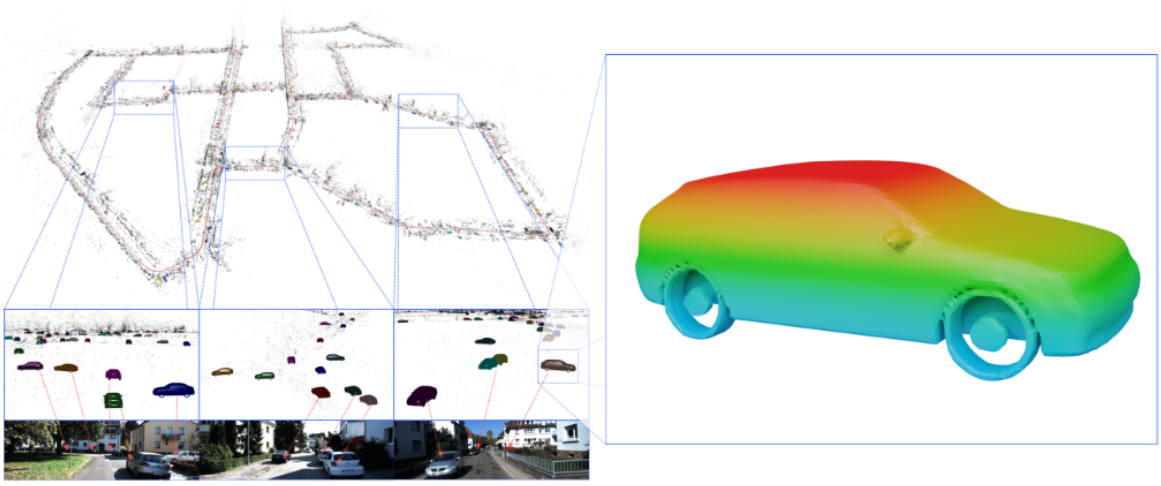


Figure 2.18: A sparse environment map, trajectory and a dense car reconstruction from KITTI 00 sequence by DSP-SLAM. Cited from Wang et al. [62].

The architecture is illustrated in Figure 2.17.

Now we conclude our survey of SLAM. The approximate timeline for the milestones of presented SLAM systems is presented in Figure 2.7, and Tables 2.2 and 2.3 recount the presented solutions in a general level respectively.

2.3 Other fields

As we have now covered photometric stereo and SLAM in the context of non-Lambertian surfaces, it is only natural to move on to the field of computer vision that is the synthesis of these two, called *fusion methods* [3]. 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. While the

Name	Approach	Density	Advancements	Drawbacks
MonoSLAM	Monocular, visual feature-based	Sparse	Monocular EKF, initialization and smooth camera movement	High memory cost
PTAM	Monocular, visual feature-based	Semi-dense	Multi-threaded tracking and mapping	Sensitive to occlusion and a lack of loop closure
DTAM	Monocular, visual direct	Dense	Data cost volume, dense mapping and accurate tracking	Lack of global optimization and loop closure
ORB-SLAM2	Feature-based	Sparse (optionally semi-dense)	Feature matching, loop closure and high optimization	Sensitive to motion blur and featurelessness
D3VO	Monocular, visual feature-based	Dense	Brightness uncertainty, increased robustness to featurelessness	Lack of loop closure

Table 2.2: A summary table of presented SLAM systems.

first study leveraging this concept dates back roughly to 1990s in the wake of shape from shading [3], the more recent related studies include exploring different penalty functions for converting the surface normals to a depth estimate by Antensteiner et al. from 2018 [3] and the usage of neural networks simultaneously to predict surface normals and depth by Zhan et al. from 2019 [72].

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 the studies of Zhang et al. [74], and Joshi et al. [29], 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 to acquire depth information lie in the problematic special

Name	Approach	Density	Advancements	Drawbacks
KinectFusion	Monocular, RGB-D	Dense	LiDAR sensing	Highly expensive in terms of memory and price, a lack of loop closure
SLAM++	Monocular, RGB-D	Semi-dense	Object tracking	Slow framerate of 20hz, requires shape priors for detection
CNN-SLAM	Monocular, visual direct	Dense	CNNs, increased robustness to featurelessness, absolute scale estimation and semantic segmentation	CPU+GPU architecture required for real-time, relies partially on baseline stereo
CubeSLAM	Monocular, visual feature-based	Sparse	Mathematical approach, info cycling and memory efficiency	Only bounding boxes for objects
DSP-SLAM	Monocular or stereo, LiDAR optional with stereo, feature-based	Dense objects, sparse environment	Detailed object reconstruction, environment maps, high optimization and run online	Slow framerate of 10hz

Table 2.3: A summary table of presented semantic SLAM systems.

cases and error margin. As pointed out by Antensteiner et al. in their survey of fusion methods in 2018 [3], a small angular error in surface normals may lead to a significant

depth error, and secondly, there is the question of penalizing the edges and planar surfaces so that the depth estimation remains accurate in both cases. One other aspect more difficult to estimate is the computational complexity driven by this approach, as the aforementioned fusion studies' is not open sourced or accurately commented on in the papers. As noted earlier in this chapter, while traditional and less costly BRDF models have been accurately fitted for a variety of materials and even general classes of materials, the neural networks so far have been most robust and general models with the expense of computation. This prompts the question whether the traditional models could be used to circumvent the computational cost of estimating the surface normals and with what disadvantages to other properties of the model.

Finally in our survey, we shall consider a multi-faceted discipline of computer vision called *illumination invariance*. The research of this topic covers diverse tasks, such as object detection [56], SLAM [32], and face recognition [77], 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. Thus while not particularly tested in the field, the used approaches may present a problem in the non-Lambertian surfaces' case as they may be devoid of geometric, texture and signal information that could otherwise be present, eg. in night- and daytime photos of same landmarks. Consequently, it would require further research whether the illumination invariant techniques would be apt to dealing with specularities, for example by quantifying the sufficient illumination for feature matching algorithms — illumination invariant ones or not — to work, while the illumination itself doesn't cause specularities to appear.

As we have now completed the survey of existing methods regarding non-Lambertian surfaces, it is only natural to look forward and what can be done next. The next chapter shall then delve into datasets regarding non-Lambertian surfaces, striving to quantify the specularities' exact impact to navigation systems and their tasks.

3. Datasets with specularities

In the preceding chapter, we covered a good deal of the existing methods to deal with specularities. Noteworthy about this survey are two very crucial questions: what are the datasets used and how good exactly are these presented methods. Indeed, these questions are even related; the answer to the latter can hardly be a definite yes, if the used datasets don't even reflect specularities properly. Thus, it is time to cast a critical look into the most common datasets, KITTI in navigation [20] and DiLiGenT in photometric stereo [53].

KITTI was first recorded in a run-of-the-mill countryside village of Karlsruhe, Germany, in 2012, primarily for the research of autonomous driving applications [20]. The dataset can be loosely described as a continuous feed of images corresponding to views a driver would see when driving leisurely around a medium-sized city and its highways on a clear sunny day. As can be noted from Figure 3.1a, specularities in a such setting are rather unequivocally minuscule, local and sparse, focused on cars and other objects comprised of common non-Lambertian surfaces, such as glass and metal. Consequently due to KITTI's popularity, SLAM systems at large are not extensively researched regarding global and dense specularities, and only three systems — CNN-SLAM [60] and D3VO [70] with their uncertainty maps and the SLAM with a novel feature matching algorithm by Dong et al. [16] — have actively paid attention to local and sparse featurelessness in their approaches to the best of author's knowledge.

Concluding, the very first obstacle when researching specularities' effect to navigation systems, one must have a new dataset, which contains not only local and sparse but global and dense specularities as well. This kind of dataset is not publicly available at the moment of writing to the best of author's knowledge, which is why I propose another dataset to be collected for this purpose. This dataset, to be published in a paper in near future, will consist of monocular video and inertial data feeds captured in an environment with an abundance of global and dense specularities: a sunny winter day with water or ice on the road. The ground truth for depth estimation will be recorded using a stereo camera and a short-range LiDAR sensor.

The second dataset to be inspected is DiLiGenT, a benchmark dataset in photometric stereo [53]. 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 [53]. The purpose is to assist the detection of specularities and to block out any noise

(a) A frame from the KITTI dataset.

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

Figure 3.1: A frame from then KITTI dataset on the left and an edited picture of the object "Lamb" from the DiLiGenT dataset on the right. As can be seen, the specularities on KITTI are local and sparse, centered around cars' corners and windows. The specularities in DiLiGenT are either similar or like in the case, soft and local.

regarding the measurement of light with finicky equipment [53], but at the same time it represents unnatural and highly ideal conditions, as can be seen in the Figure 3.1b. Furthermore, even the specularities in DiLiGenT are small and mostly local and sparse in nature, thus not accounting for global and dense specularities properly. Thus, even in the field of photometric stereo, a more challenging benchmark dataset accounting for global and dense specularities in objects could be introduced.

Thus, the next section shall focus on a new dataset collected by the author. While not publicly available, it can be requested from the author via email. The task the dataset is most suited for is single-view monocular depth estimation — that is, a subfield of navigation, where a dense depth is estimated from a single image and specularities hold potential for complications, as there is no prior or subsequent info available. In the setting, a special attention was given to challenging lighting conditions, ranging from darkness to overbrightness.

3.1 Single-view monocular depth estimation

3.1.1 Data collection

In this dataset collection, the aim was to collect data, which could be used to to quantify the effect of specularities in single-view monocular depth estimation. The data consists of ground truth depth data and monocular RGB images used as a test set. The used equipment was a RealSense D435i stereo depth camera and a GoPro Max camera, whose details are presented on Table 3.1. The cameras were positioned on a

RealSense D435i	
Width	424px
Height	240px
Bytes per pixel	2
F_x	213.326874
F_y	213.326874
PP_x	211.939163
PP_y	120.154716
Distortion	Brown Conrady
Frame format	Z16
Frame type	Depth
File format	.PNG, .RAW

(a) A table about RealSense D435i camera calibration parameters for the first ground truth picture of natural illumination test set.

GoPro Max	
Width	2704px
Height	2028px
Resolution	4MP
Focal length	3.00 mm
Shutter	1/156 s
F-value	2.8
ISO	300
White balance	Auto
Flash	None
Frame type	RGB
File format	.JPG

(b) A table about GoPro Max camera parameters for the natural illumination test set.

Table 3.1: Examples of parameters used in the RealSense and GoPro Max cameras during the dataset collection.

book pile atop an office table, located in a private office room at Exactum, a building in University of Helsinki's Kumpula campus, and angled towards a window and various objects on the office table, such as a desk lamp, an award plate and a decorative picture. Two subsets of data were gathered on both cameras, one set consisting of images where the desk lamp on the table was turned on, referred to as "the manual illumination" in Table 3.2, and later, another set where the lamp was off, referred to as "the natural illumination" in Table 3.2. A sample of each subset is presented in Figure 3.3.

The nature of the second dataset should be brought to the reader's attention. The GoPro Max camera was set to capture one frame each minute on its own until stopped. This continuous timelapse was run a little over 18 hours, totaling 1090 pictures. Some of the notable illumination changes captured during the timelapse — referred to as "the natural illumination" — include the gradual day and night cycle, as well as a powerful and bright reflection coming from the window, possibly caused by overexposure or a specularity angled right towards the camera. In comparison, the first set of pictures collected with the GoPro Max camera with the desk lamp turned on — referred to as "the manual illumination" — include stronger and larger specularities in the award plate and near the window, and one softer specularity on the desk. The exact sizes and collection times of each test and ground truth dataset are detailed in Table 3.2.

However, as many of the timelapse pictures exhibit barely perceivable differences

Dataset	Camera	Partition	Date	Time	Size
Ground truth	RealSense D435i	Manual illumination	21.6.2022	15.20-15.30	7 pictures
		Natural illumination	21.6.2022	15.40-15.50	5 pictures
Test set	GoPro Max	Manual illumination	21.6.2022	15.30.-15.45	13 pictures
		Natural illumination	21.-22.6.2022	15.52-10.02	1090 pictures

Table 3.2: The table describing the datasets collected.

in short time intervals, a smaller dataset with all the notable phenomena visible was chosen for the tests. This set contains 19 pictures from the natural illumination test set, taken on the hour between 16.00 and 10.00, and the fifth picture from the manual illumination test set. The fifth picture from the natural illumination ground truth data set, having more pixels than the manual illumination ground truth pictures, was chosen for the evaluation.

3.1.2 Results

The testing of monocular depth estimation was carried out using DNet by Xue et al. [69]. The reasons for the choice were numerous: modern state-of-the-art performance in KITTI Eigen Split, the usage of surface normals in the algorithm, the estimation of absolute depth for effortless evaluation against RealSense data, and the availability of a public demo notebook for non-commercial uses. DNet's novelties revolve around a dense connected prediction (DCP) layer and the estimation of camera height. The DCP layer can combine features from multiple scales leading to more accurate object-level depths and boundaries. Then, assuming dense ground level points and using surface normals to estimate the camera height, DNet solves the scaling factor required in monocular SLAM systems to obtain the absolute depth estimation from the relative depth estimation by dividing the true camera height with the estimated camera height.

The metrics of running DNet against the mini test set with the pre-trained model from the paper of Xue et al. are presented in Table 3.3, while the visualized depth estimations are presented in Figure 3.7. As can be deduced from the Table 3.3 and Figure 3.7, DNet's accuracy is fairly good, while not outstanding: the depth error ranges roughly from 15 centimeters to 60, with incorrect predictions focusing around the window and the wall on the background. Furthermore, the performance suffers most

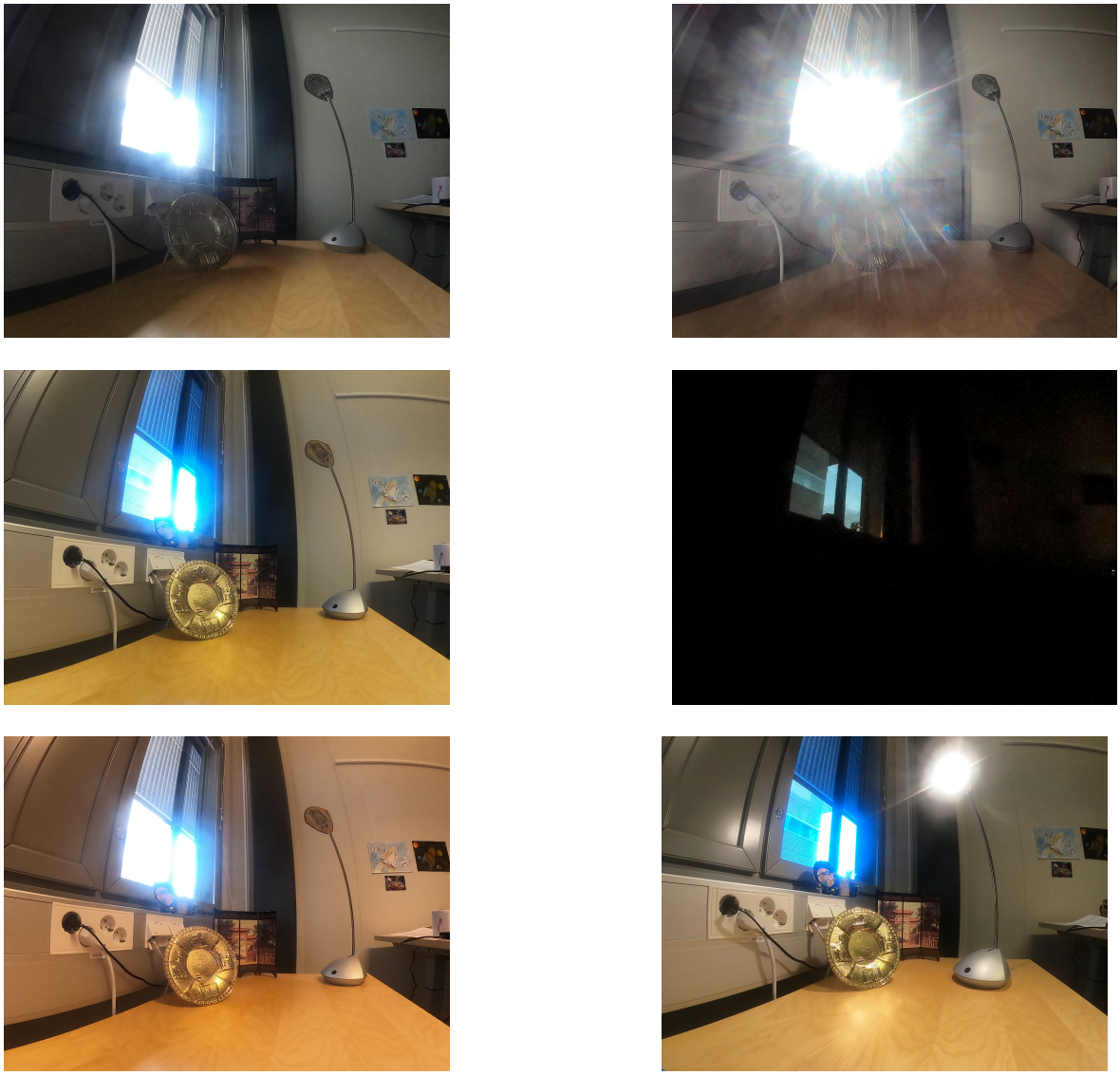
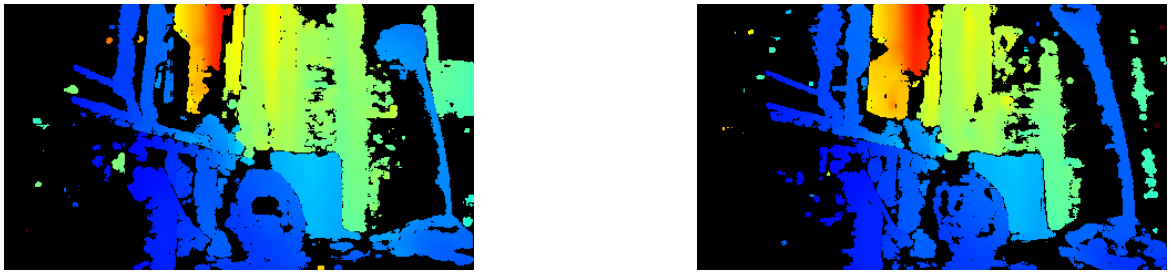


Figure 3.2: Samples from the collected test datasets. The picture on the right lower corner is from the manually enhanced illumination set, while others are from the natural illumination set.



(a) The natural illumination ground truth.

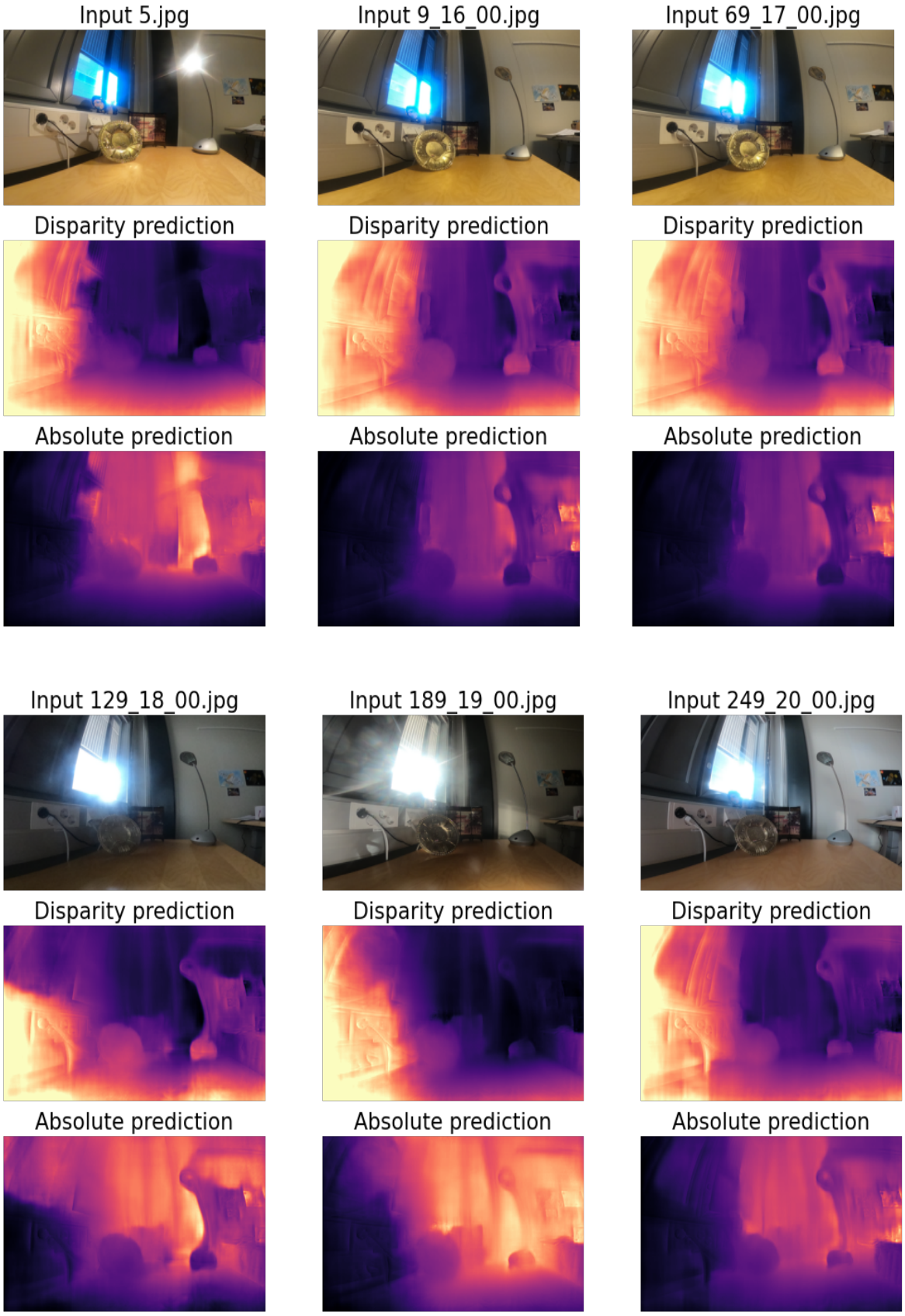
(b) The manual illumination ground truth.

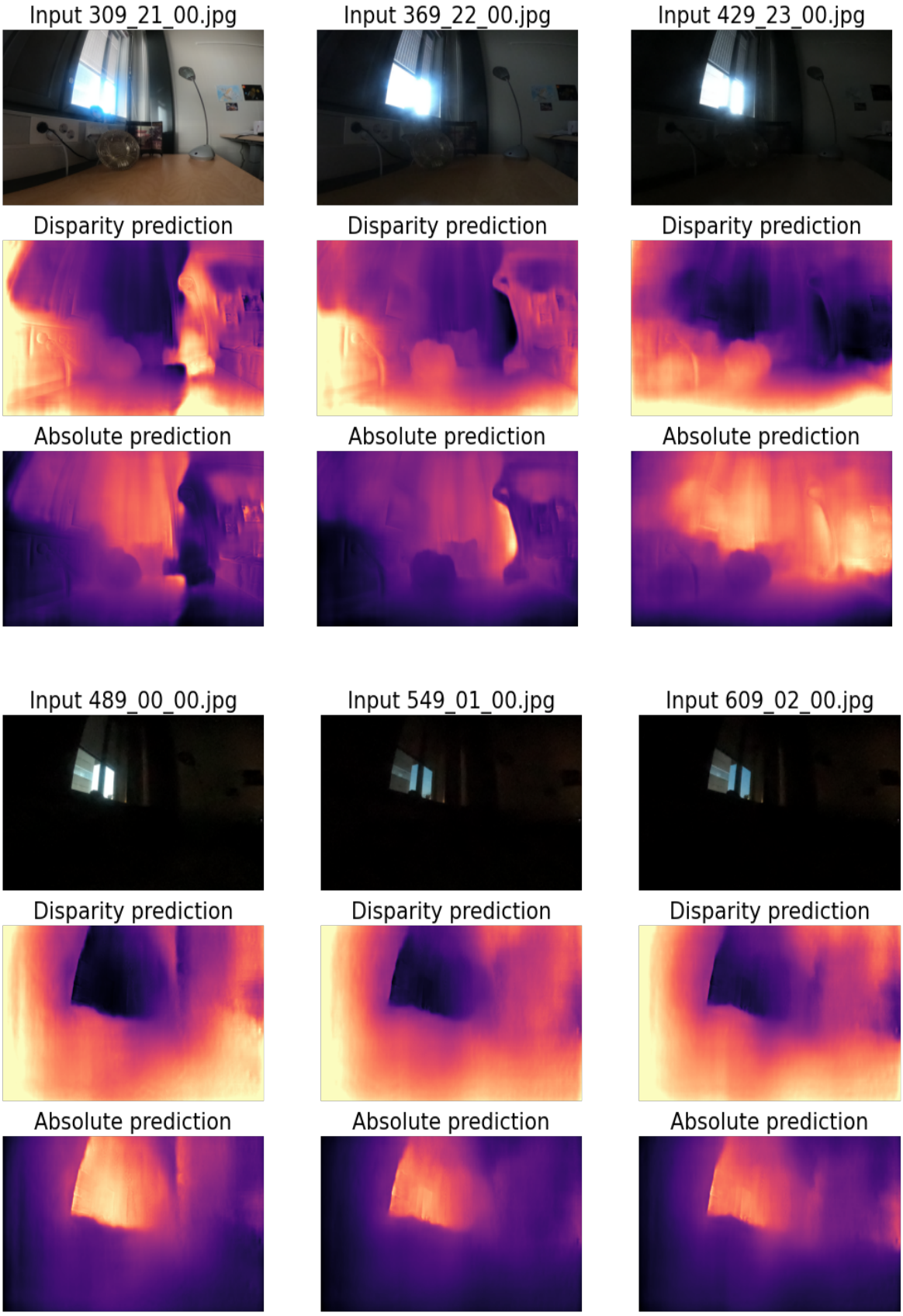
Figure 3.3: Samples from the collected ground truth datasets.

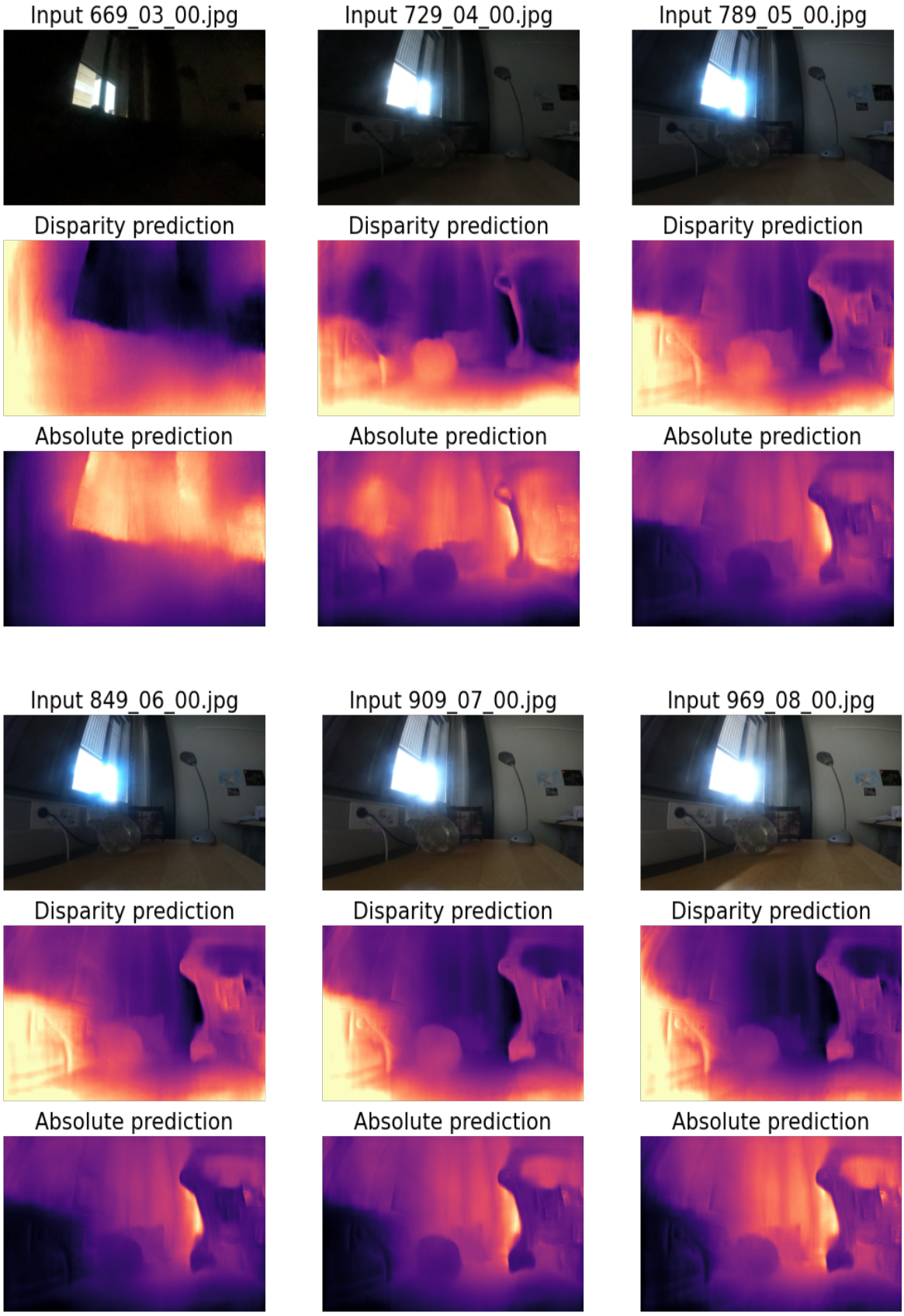
Picture	Abs.rel.	Sq.rel.	RMSE	Log RMSE	1.25^1	1.25^2	1.25^3
9_16_00.jpg	0.314	0.124	0.382	0.401	0.345	0.749	0.941
69_17_00.jpg	0.334	0.132	0.380	0.402	0.331	0.718	0.933
129_18_00.jpg	0.363	0.123	0.327	0.381	0.328	0.728	0.958
189_19_00.jpg	0.268	0.112	0.387	0.386	0.398	0.740	0.937
249_20_00.jpg	0.280	0.100	0.350	0.358	0.353	0.826	0.953
309_21_00.jpg	0.409	0.159	0.378	0.466	0.300	0.569	0.857
369_22_00.jpg	0.366	0.140	0.392	0.414	0.250	0.707	0.901
429_23_00.jpg	0.370	0.140	0.376	0.410	0.256	0.690	0.922
489_00_00.jpg	0.421	0.172	0.347	0.426	0.311	0.697	0.887
549_01_00.jpg	0.373	0.160	0.401	0.448	0.287	0.560	0.899
609_02_00.jpg	0.362	0.150	0.400	0.438	0.270	0.594	0.900
669_03_00.jpg	0.322	0.107	0.319	0.356	0.343	0.783	0.976
729_04_00.jpg	0.384	0.149	0.371	0.415	0.261	0.711	0.909
789_05_00.jpg	0.349	0.127	0.364	0.396	0.275	0.755	0.919
849_06_00.jpg	0.338	0.119	0.357	0.387	0.285	0.746	0.957
909_07_00.jpg	0.395	0.147	0.382	0.417	0.252	0.646	0.944
969_08_00.jpg	0.360	0.126	0.354	0.398	0.275	0.687	0.966
1029_09_00.jpg	0.331	0.117	0.365	0.389	0.267	0.735	0.956
1089_10_00.jpg	0.323	0.128	0.376	0.391	0.352	0.747	0.937
5	0.306	0.107	0.312	0.346	0.491	0.772	0.967

Table 3.3: Results of the depth estimation experiment against the mini test set.

from medium darkness accompanied with bright light than from the total darkness or soft specularities. Thus, it can reasonably be concluded that DNet is passable as an out-of-the-box solution in settings similar to the mini test test, and could be trained with data similar to the mini test set if there is a need to increase the performance. The sensor rig could additionally be equipped with a shadow or light source of its own, to mitigate the effect of mixed illumination as needed. KITTI's characteristics support these notions: it is recorded in broad daylight outdoors and has a wildly different scale of objects and distances than our simple setting indoors. Summarizing, it can be concluded that extra care must be put into the training of DNet so that it may cope with different aspects of featurelessness. The definitive answer whether this holds for alternative SLAM systems would require further studies, but the results of DNet indicate a need for it the very least. Settings with global and dense specularities would be good to investigate as well.


Input 5.jpg
Input 9_16_00.jpg
Input 69_17_00.jpg
Disparity prediction
Disparity prediction
Disparity prediction
Absolute prediction
Absolute prediction
Absolute prediction
Input 129_18_00.jpg
Input 189_19_00.jpg
Input 249_20_00.jpg
Disparity prediction
Disparity prediction
Disparity prediction
Absolute prediction
Absolute prediction
Absolute prediction


Input 309_21_00.jpg
Input 369_22_00.jpg
Input 429_23_00.jpg
Disparity prediction
Disparity prediction
Disparity prediction
Absolute prediction
Absolute prediction
Absolute prediction
Input 489_00_00.jpg
Input 549_01_00.jpg
Input 609_02_00.jpg
Disparity prediction
Disparity prediction
Disparity prediction
Absolute prediction
Absolute prediction
Absolute prediction


Input 669_03_00.jpg
Input 729_04_00.jpg
Input 789_05_00.jpg
Disparity prediction
Disparity prediction
Disparity prediction
Absolute prediction
Absolute prediction
Absolute prediction
Input 849_06_00.jpg
Input 909_07_00.jpg
Input 969_08_00.jpg
Disparity prediction
Disparity prediction
Disparity prediction
Absolute prediction
Absolute prediction
Absolute prediction

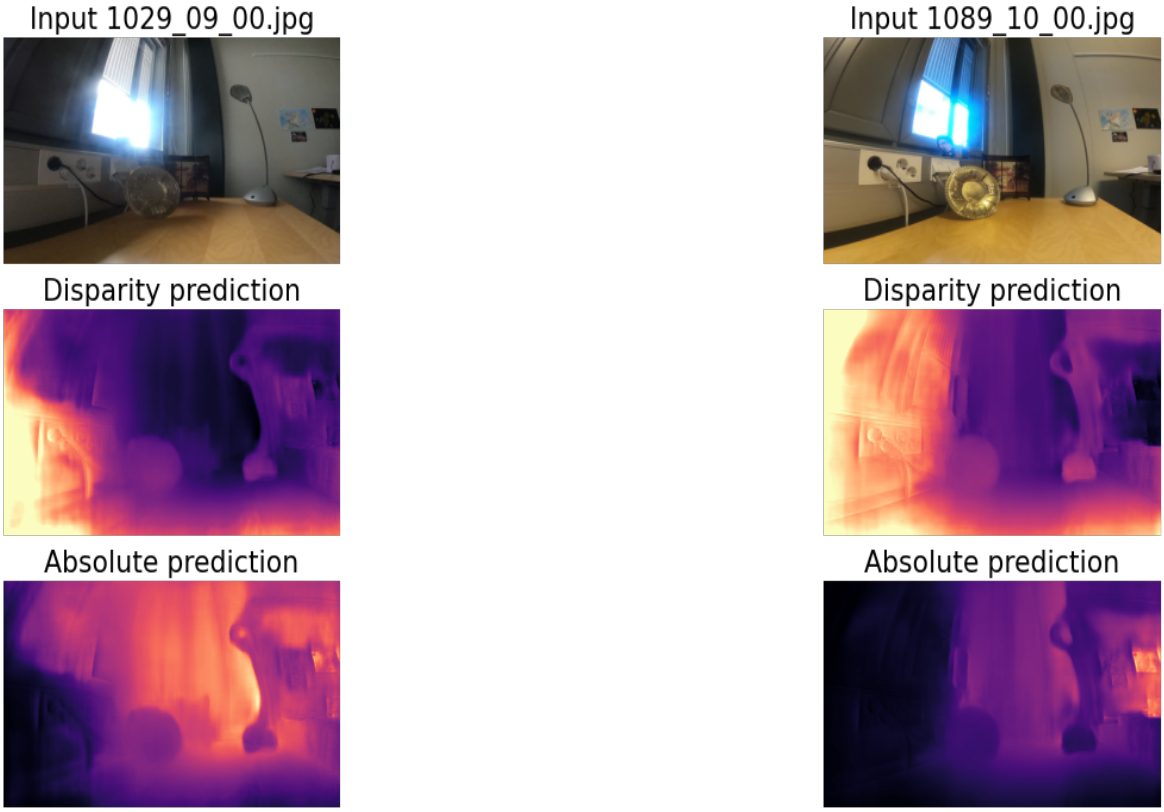


Figure 3.7: The visualized depth estimations of DNet running against the mini test set. In disparity predictions (the middle images), light coloured areas have most disparity and the dark areas the least. In absolute depth predictions (the lowest images), the closest areas are dark and the areas farther away are bright. It can be clearly observed that the changing illumination clearly impacts the performance and depth.

4. Discussion

As of now, we have gone through a wide range of studies regarding non-Lambertian surfaces. Quickly recounting, from photometric stereo there are

- outlier methods, which deal well with bright, local and sparse specularities, but have often difficulties with soft, global and dense ones
- analytical BRDFs, which suffer from computational complexity and poor generalization
- masking methods, which have limiting assumptions, as the relevancy of nearby pixels or a uniform distribution of colours
- generalized BRDFs and bi-polynomial approximations, which fail with the special cases violating the assumptions and finally
- neural networks, which also fail with the special cases, have poor explainability and dismiss the BRDF altogether.

As for navigation, D3VO and CNN-SLAM have utilized varying uncertainty mappings to refine the depth estimation in case of specularities, whereas semantic SLAMs, such as DSP-SLAM, and the SLAM system by Dong et al., have been able to use objects to reconstruct info underneath the specularities. However, there is little or no data available regarding these methods' performance against global and dense specularities as far as I am aware, and thus new studies about that should be conducted. Another noteworthy point to be solved is how each system's specific drawbacks might affect the navigation tasks under global and dense specularities.

Summarizing, the most viable source of information we can gain from specularities is the geometric and reflection information. As the specularity is caused by a reflection from a surface, both from the reflection and the surface can be deduced several facts, such as the approximate position of the surface and the characteristic reflection of the surface. While this information can come with a measure of uncertainty, it still holds potential for future, and sincerely I can say that I expect most benefit to come from the union of the models of photometric stereo, the optimized neural network computations and system architectures of SLAM, and the uncertainty or shape mappings. Fusion methods especially hold potential for overcoming the specularities, given that the entailed computational complexity is not too much. Boldly assuming this, a new kind of approach could be envisioned, which I refer to as the *reflection segmentation*. In this

approach, different objects and materials could be classified with different reflection profiles, such as the Ward model, the bi-polynomial approximation and so on, using an uncertainty mapping over the mismatch of the data and the implied parameters to choose the best fitting reflection model. Using this chosen model, the surface normals of the object or surface could be estimated and fed into a pipeline that converts the surface normals into absolute depth estimation, much like in the study of Antensteiner et al. [3]. As considerable accuracy was already achieved in their study — by average with a mean square error between 0.1 and 0.3 with the method of generalized Nehab [3] — and the same level of accuracy would be sufficient for navigation, this whole pipeline of surface normals and depth estimation might be the future of SLAM especially. Indeed, surface normals can be predicted at the same time as the depth with the neural networks, as proven by Eigen et al. [19], Weerasekera et al. [66] and Zue et al. [69] in their studies from 2014, 2017 and 2020 respectively. However, to do a such pipeline in practice, a suitable dataset for the reflection segmentation testing should be acquired. While the dataset presented in Chapter 3 is not straightforwardly meeting this goal, it could still be used with semantic labelling and the DiLiGenT dataset is viable as well and better equipped due to its larger size, as hundreds of images may be required for anisotropic reflection and outdoor settings among other tasks.

The problem of suitable datasets regarding the study of non-Lambertian surfaces, particularly in multidisciplinary topics of computer vision, is indeed the main hurdle for now as was noted in the previous chapter. The collected dataset is nevertheless far from enough, and it has its own shortcomings, like softer lighting and an absence of global and dense specularities. It does, however, address a larger target audience of illumination invariance research as well and may reveal other useful information besides the system's accuracy, such as the most useful lighting conditions for navigation. The outlined dataset may well be even more generalizable to other tasks, such as object detection and reconstruction, and thus has unique potential for future research of specularities. Another noteworthy research topic in SLAM would be a new formulation of photometric error that doesn't rely on brightness constancy and which could be used as a metric even with specularities present in the data.

As a concluding remark, we can still note that the elemental issues behind specularities — the violations of brightness and colour constancies — are still there and have noteworthy potential to curb the existing methods' efficiency with global and dense specularities. Without empirical studies, it can be only hypothesized whether a shape prior can be sufficiently deduced for object reconstruction, whether a less bright picture of non-Lambertian surface obtained for depth estimation or whether the assumptions of masking methods met sufficiently in practical settings. Thus I would call for more research around this topic to fully determine what significance non-Lambertian surfaces

can hold, and whether the environment can work to our advantage or disadvantage regarding them, as is investigated in illumination invariance.

5. Conclusions

This thesis has focused on non-Lambertian surfaces in the field of computer vision, casting an in-depth review into their fundamental theory, issued posed for various tasks, their respective solutions and finally the future steps for research.

The thesis was opened by presenting the elemental optics and the two frameworks required for the modelling of non-Lambertian surfaces: the Lambertian reflectance model and the bi-directional reflectance distribution function. The first defines Lambertian surfaces as ideally diffuse surfaces, whose luminance is isotropic and the luminous intensity obeys Lambert's cosine law, and non-Lambertian surfaces as the opposite of Lambertian surfaces, capable of creating specularities and violating either of the aforementioned assumptions. This knowledge then gave a starting point for the more general model: the bi-directional reflectance distribution function, shortened as the BRDF, where the reflected light is a function of the incident light and surface normal. By further investigating the definition of non-Lambertian surfaces, we concluded that they contradict the brightness and colour constancies frequently used as assumptions in computer vision algorithms and thus can cause an abundance of complications. An alternative and slightly exaggerating phrasing of the issue is how we can say that two camera coordinates are the same real-world coordinates, if the direct info available from the coordinates is not the same.

Using the BRDF, a survey into photometric stereo could be conducted. Recounting historically significant studies and theoretical advancements, along with modern state-of-the-art methods with an extensive overview of pros and cons of existing methods was presented. As examples of the former kind of studies are Woodham [67], Hayakawa [24] and Rusienkiewicz [49], and from the latter PS-FCN [8], LCNet [6] and the inverse reflectance model from Wang et al. [30].

After photometric stereo the survey's focus was shifted into another field of computer vision, where specularities have been known to cause complications: navigation. Presenting the probabilistic formulation of Simultaneous Localization and Mapping (SLAM), various important systems involving monocular cameras were presented along with SLAM systems capable of handling specularities in navigation, such as MonoSLAM [14] and ORB-SLAM2 [41] for the former and DSP-SLAM [62] and D3VO [70] for the latter. Finally a brief account of other potential fields — namely fusion methods and illumination invariance — for the research of non-Lambertian surfaces was given.

Concluding from the survey, the shortcomings of common benchmark datasets

used in photometric stereo and navigation were pointed out regarding the global and dense specularities, in contrast to extensively studied local and sparse specularities. A new dataset including extreme lighting conditions and soft specularities was then collected and presented, as to provide an alternative dataset from natural conditions with larger specularities to the purpose of single-view monocular depth estimation. Another dataset to be later released in a paper, aimed for the tasks of tracking and mapping under global and dense specularities, was designed and generally outlined.

Finally, a critical discussion into the presented methods and collected dataset was underwent and a theoretical description regarding a potential algorithm was given. The main challenges of non-Lambertian surfaces and their research were summarized and reflected on.

Concluding from all of this, it can be said that non-Lambertian surfaces have not been given the scientific attention in computer vision they are deserving. As they contradict the common assumptions of brightness and colour constancy, new tools to handle them adequately and generally are required, which in turn expect a plenitude of research. Especially global and dense specularities remain as a negligible subject, while imposing severe theoretical consequences for the existing methods. The first necessary step for the discovery of truly generalized and efficient solutions is to acquire datasets introducing diverse specularities, and preferably numerous non-Lambertian surfaces. Only after this, can a considerable effort be given to the testing of exiting methods and the development of suitable new algorithms, ranging from photometric stereo and navigation to feature matching algorithms and object tracking.

Bibliography

[1] J. Ackermann, F. Langguth, S. Fuhrmann, and M. Goesele. Photometric stereo for outdoor webcams. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 262–269, 2012.

[2] N. G. Alldrin, S. P. Mallick, and D. J. Kriegman. Resolving the Generalized Bas-Relief Ambiguity by Entropy Minimization. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7, 2007.

[3] D. Antensteiner, S. Štolc, and T. Pock. A Review of Depth and Normal Fusion Algorithms. *Sensors (Basel)*, 18:431, 2018.

[4] T. Bailey and H. Durrant-Whyte. Simultaneous localization and mapping (slam): part ii. *IEEE Robotics Automation Magazine*, 13(3):108–117, 2006.

[5] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard. Past, Present, and Future of Simultaneous Localization and Mapping: Toward the Robust-Perception Age. *IEEE Transactions on Robotics*, 32(6):1309–1332, 2016.

[6] G. Chen, K. Han, B. Shi, Y. Matsushita, and K.-Y. K. Wong. Deep Photometric Stereo for Non-Lambertian Surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP:1–1, 2020.

[7] G. Chen, K. Han, B. Shi, Y. Matsushita, and K.-Y. K. K. Wong. Self-calibrating deep photometric stereo networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8731–8739, 2019.

[8] G. Chen, K. Han, and K.-Y. K. Wong. PS-FCN: A Flexible Learning Framework for Photometric Stereo, 2018.

[9] F. Chhaya, D. Reddy, S. Upadhyay, V. Chari, M. Z. Zia, and K. M. A. Krishna. Monocular reconstruction of vehicles: Combining SLAM with shape priors. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5758–5765, 2016.

[10] H.-S. Chung and J. Jia. Efficient photometric stereo on glossy surfaces with wide specular lobes. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[11] E. N. Coleman and R. Jain. Obtaining 3-dimensional shape of textured and specular surfaces using four-source photometry. *Computer Graphics and Image Processing*, 18(4):309–328, 1982.

[12] A. Dame, V. A. Prisacariu, C. Y. Ren, and I. Reid. Dense reconstruction using 3d object shape priors. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1288–1295, 2013.

[13] J. P. David A. Forsyth. *Computer vision: A modern approach*. Prentice Hall, 2 edition, 2011.

[14] A. Davison, I. Reid, N. Molton, and O. Stasse. MonoSLAM: real-time single camera SLAM. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29:1052–67, 07 2007.

[15] A. Del Bue, J. Xavier, L. Agapito, and M. Paladini. Bilinear Factorization via Augmented Lagrange Multipliers. In *ECCV (4)*, pages 283–296, 09 2010.

[16] Y. Dong, S. Wang, J. Yue, C. Chen, S. He, H. Wang, and B. He. A Novel Texture-Less Object Oriented Visual SLAM System. *IEEE Transactions on Intelligent Transportation Systems*, 22(1):36–49, 2021.

[17] H. Durrant-Whyte and T. Bailey. Simultaneous localization and mapping: part i. *IEEE Robotics Automation Magazine*, 13(2):99–110, 2006.

[18] A. Earp, G. Smith, and J. Franklin. Simplified BRDF of a Non-Lambertian Diffuse Surface. *Lighting research & technology (London, England : 2001)*, 39(3):265–281, 2007.

[19] D. Eigen and R. Fergus. Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-Scale Convolutional Architecture, 2014.

[20] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012.

[21] Georghiades. Incorporating the Torrance and Sparrow model of reflectance in uncalibrated photometric stereo. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 816–823 vol.2, 2003.

[22] D. B. Goldman, B. Curless, A. Hertzmann, and S. M. Seitz. Shape and Spatially-Varying BRDFs from Photometric Stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(6):1060–1071, 2010.

[23] S. M. Haque, A. Chatterjee, and V. M. Govindu. High quality photometric reconstruction using a depth camera. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2283–2290, 2014.

[24] H. Hayakawa. Photometric stereo under a light source with arbitrary motion. *Journal of The Optical Society of America A-optics Image Science and Vision*, 11:3079–3089, 1994.

[25] S. Hoseini and P. Kabiri. A Novel Feature-Based Approach for Indoor Monocular SLAM. *Electronics (Basel)*, 7(11):305–320, 2018.

[26] C. Häne, N. Savinov, and M. Pollefeys. Class Specific 3D Object Shape Priors Using Surface Normals. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 652–659, 2014.

[27] S. Ikehata, D. Wipf, Y. Matsushita, and K. Aizawa. Robust photometric stereo using sparse regression. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 318–325, 2012.

[28] V. Ila, L. Polok, M. Solony, and P. Svoboda. Slam++ -a highly efficient and temporally scalable incremental slam framework. *The International Journal of Robotics Research*, 36:027836491769111, 02 2017.

[29] N. Joshi and D. J. Kriegman. Shape from varying illumination and viewpoint. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–7, 2007.

[30] B. Kaya, S. Kumar, C. Oliveira, V. Ferrari, and L. Van Gool. Uncalibrated Neural Inverse Rendering for Photometric Stereo of General Surfaces. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 06 2021.

[31] G. Klein and D. Murray. Parallel tracking and mapping for small ar workspaces. In *2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*, pages 225–234, 2007.

[32] M. Labbé and F. Michaud. Multi-Session Visual SLAM for Illumination-Invariant Re-Localization in Indoor Environments. *Frontiers in Robotics and AI*, 9, jun 2022.

[33] K. Li, M. Rünz, M. Tang, L. Ma, C. Kong, T. Schmidt, I. Reid, L. Agapito, J. Straub, S. Lovegrove, and R. Newcombe. FroDO: From Detections to 3D Objects. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[34] J. Lim, J. Ho, M.-H. Yang, and D. Kriegman. Passive photometric stereo from motion. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pages 1635–1642 Vol. 2, 2005.

[35] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015.

[36] F. Lu, I. Sato, and Y. Sato. Uncalibrated photometric stereo based on elevation angle recovery from BRDF symmetry of isotropic materials. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 168–176, 2015.

[37] W. R. McCluney. *Introduction to radiometry and photometry.* The Artech House optoelectronics library. Artech House, 2 edition, 2014.

[38] O. Meslouhi, M. Kardouchi, H. Allali, T. Gadi, and Y. Benkaddour. Automatic detection and inpainting of specular reflections for colposcopic images. *Open computer science*, 1(3):341–354, 2011.

[39] A. Mirko, G. Anarta, A. Stefan, and G. Lacey. Automatic Segmentation and Inpainting of Specular Highlights for Endoscopic Imaging. *EURASIP Journal on Image and Video Processing*, 2010(9), 01 2010.

[40] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós. ORB-SLAM: A Versatile and Accurate Monocular SLAM System. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015.

[41] R. Mur-Artal and J. D. Tardós. ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras. *IEEE Transactions on Robotics*, 33(5):1255–1262, 2017.

[42] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon. KinectFusion: Real-time dense surface mapping and tracking. In *2011 10th IEEE International Symposium on Mixed and Augmented Reality*, pages 127–136, 2011.

[43] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison. DTAM: Dense tracking and mapping in real-time. In *2011 International Conference on Computer Vision*, pages 2320–2327, 2011.

[44] T. Papadhimitri and P. Favaro. A Closed-Form, Consistent and Robust Solution to Uncalibrated Photometric Stereo Via Local Diffuse Reflectance Maxima. *International Journal of Computer Vision*, 107, 04 2014.

[45] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove. DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 165–174, 2019.

[46] M. Pharr, W. Jakob, and G. Humphreys. *Physically Based Rendering: From Theory To Implementation*. Published online: https://www.pbr-book.org/3ed-2018/contents, 3rd edition, 2018. Accessed September 6, 2022.

[47] S. L. Richard Newcombe and A. Davison. Dtam: Dense tracking and mapping in real-timen, 2011. Accessed online September 6, 2022.

[48] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. ORB: an efficient alternative to SIFT or SURF. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2564–2571, 11 2011.

[49] S. M. Rusinkiewicz. A New Change of Variables for Efficient BRDF Representation. In G. Drettakis and N. Max, editors, *Rendering Techniques '98*, pages 11–22, Vienna, 1998. Springer Vienna.

[50] 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. *Machine vision and applications*, 22(1):171–180, 2007.

[51] I. Sato, T. Okabe, Q. Yu, and Y. Sato. Shape Reconstruction Based on Similarity in Radiance Changes under Varying Illumination. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8, 2007.

[52] B. Shi, Y. Matsushita, Y. Wei, C. Xu, and P. Tan. Self-calibrating photometric stereo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1118–1125, 2010.

[53] 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 Transactions on Pattern Analysis and Machine Intelligence*, 41(2):271–284, 2019.

[54] B. Shi, P. Tan, Y. Matsushita, and K. Ikeuchi. Bi-Polynomial Modeling of Low-Frequency Reflectances. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(6):1078–1091, 2014.

[55] Simakov, Frolova, and Basri. Dense shape reconstruction of a moving object under arbitrary, unknown lighting. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 1202–1209 vol.2, 2003.

[56] D. Slater and G. Healey. The illumination-invariant recognition of 3d objects using local color invariants. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(2):206–210, 1996.

[57] E. Sucar, K. Wada, and A. Davison. NodeSLAM: Neural object descriptors for multi-view shape reconstruction. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2020.

[58] R. Szeliski. Computer Vision Algorithms and Applications. In *Computer Vision Algorithms and Applications*, Texts in Computer Science, pages 220–231. Springer London, London, 2nd ed. 2021. edition, 2021.

[59] T. Taniai and T. Maehara. Neural Inverse Rendering for General Reflectance Photometric Stereo, 2018.

[60] K. Tateno, F. Tombari, I. Laina, and N. Navab. Cnn-slam: Real-time dense monocular slam with learned depth prediction, 2017.

[61] F. Verbiest and L. Van Gool. Photometric stereo with coherent outlier handling and confidence estimation. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[62] J. Wang, M. Rünz, and L. Agapito. DSP-SLAM: Object Oriented SLAM with Deep Shape Priors. In *2021 International Conference on 3D Vision (3DV)*, pages 1362–1371, 2021.

[63] 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*, 29:6032–6042, 2020.

[64] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[65] G. J. Ward. Measuring and modeling anisotropic reflection. In *Proceedings of the 19th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '92, page 265–272, New York, NY, USA, 1992. Association for Computing Machinery.

[66] C. S. Weerasekera, Y. Latif, R. Garg, and I. Reid. Dense monocular reconstruction using surface normals. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2524–2531, 2017.

[67] R. J. Woodham. Photometric Method For Determining Surface Orientation From Multiple Images. *Optical Engineering*, 19(1):139 – 144, 1980.

[68] T.-P. Wu and C.-K. Tang. Photometric stereo via expectation maximization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32:546 – 560, 04 2010.

[69] F. Xue, G. Zhuo, Z. Huang, W. Fu, Z. Wu, and M. H. Ang. Toward hierarchical self-supervised monocular absolute depth estimation for autonomous driving applications. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2330–2337. IEEE, 2020.

[70] N. Yang, L. von Stumberg, R. Wang, and D. Cremers. D3VO: Deep Depth, Deep Pose and Deep Uncertainty for Monocular Visual Odometry. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[71] S. Yang and S. Scherer. CubeSLAM: Monocular 3-D Object SLAM. *IEEE Transactions on Robotics*, 35(4):925–938, Aug 2019.

[72] H. Zhan, C. S. Weerasekera, R. Garg, and I. Reid. Self-supervised learning for single view depth and surface normal estimation, 2019.

[73] G. Zhang and P. A. Vela. Good features to track for visual SLAM. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1373–1382, 2015.

[74] L. Zhang, Curless, Hertzmann, and Seitz. Shape and motion under varying illumination: unifying structure from motion, photometric stereo, and multiview stereo. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 618–625 vol.1, 2003.

[75] Q. Zhang, M. Ye, R. Yang, Y. Matsushita, B. Wilburn, and H. Yu. Edge-preserving photometric stereo via depth fusion. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2472–2479, 2012.

[76] 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 Transactions on Pattern Analysis and Machine Intelligence*, 42(7):1594–1605, 2020.

[77] X. Zou, J. Kittler, and K. Messer. Illumination Invariant Face Recognition: A Survey. In *2007 First IEEE International Conference on Biometrics: Theory, Applications, and Systems*, pages 1–8, 2007.