Technische Universität München

TUM School of Computation, Information and Technology

# Physically-Based Inverse Problems for High-Quality 3D Reconstruction

Björn Häfner

Vollständiger Abdruck der von der TUM School of Computation, Information and Technology der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften
(Dr. rer. nat.)

genehmigten Dissertation.

| | |
|---|---|
| Vorsitz: | Prof. Dr. Rüdiger Westermann |
| Prüfer der Dissertation: | 1. Prof. Dr. Daniel Cremers |
| | 2. Prof. Dr. Bastian Goldlücke |
| | 3. Prof. Dr. Ronen Basri |

Die Dissertation wurde am 24.07.2023 bei der Technischen Universität München eingereicht und durch die TUM School of Computation, Information and Technology am 11.11.2024 angenommen.

To my family.
*– For all the love and support.*

To Emilia and Dominik.
*They wished their daddy wrote a children's book.*
*– I love you dearly.*

*If you can't explain it simply, you don't understand it well enough.*

Albert Einstein (1879 – 1955)

# Abstract

A fundamental research area in computer vision and computer graphics has been 3D reconstruction, aiming to model the relationship between the real world and its corresponding images. However, many existing approaches encounter limitations in accurately modeling hardware characteristics, assuming controlled laboratory environments, or being applicable only to single-object scenes. This thesis comprises four papers that specifically tackle these challenges in the area of physically-based 3D reconstruction.
First, we propose a principled variational approach for up-sampling a single depth map to match the resolution of a companion color image from an RGB-D sensor. By combining depth and color data, we simultaneously address the depth super-resolution (SR) and Shape-from-Shading (SfS) problems. By accurately modeling the discrepancy in resolution of RGB-D sensors, we demonstrate that the extraction of low-frequency geometric information from low-resolution (LR) depth maps can disambiguate SfS and that the use of high-frequency photometric clues from the RGB image can disambiguate depth SR.
Second, we build on the principles of photometric stereo (PS) where we introduce an efficient variational approach to uncalibrated PS under general illumination conditions. We approximate the Lambertian reflectance model through a spherical harmonic expansion, enabling the recovery of shape, reflectance, and illumination as a single variational problem. The proposed approach utilizes an innovative minimal surface initialization and additionally eliminates the need for subsequent normal integration.
Next, we address the joint problem of 3D reconstruction and 2D segmentation of objects using PS. Unlike previous works, which assume a precomputed mask of the area of interest, we formulate the reconstruction and segmentation as a joint problem. By combining a differential formulation of PS with the Chan-Vese model for active contours, the proposed variational solution simultaneously infers a binary mask of the object of interest and a depth map.
Finally, we present a method to estimate the bidirectional reflectance distribution function (BRDF) of complete scenes in an uncontrolled environment. We split the BRDF into diffuse and non-diffuse components and solve for each component separately. By employing Monte Carlo ray tracing and a ray-tracing-based optimization strategy, we efficiently estimate all parameters of the BRDF.
Overall, we address critical problems in physically-based 3D reconstruction such as depth SR, SfS, uncalibrated PS, joint PS and segmentation, as well as BRDF estimation. Evaluated on challenging synthetic and real-world data, the proposed approaches demonstrate improved performance, efficiency, and robustness compared to existing methods, paving the way for further developments and applications in the field.

# Zusammenfassung

Ein bedeutender Forschungsbereich in der Computer Vision und Computergrafik ist die 3D-Rekonstruktion, die darauf abzielt, die Beziehung zwischen der realen Welt und den entsprechenden Bildern zu modellieren. Viele bestehende Ansätze stoßen jedoch an ihre Grenzen, wenn es darum geht, Hardwareeigenschaften genau zu modellieren, kontrollierte Laborumgebungen zu relaxieren oder auf mehr als nur auf Einzelobjektszenen anwendbar zu sein. Diese Arbeit umfasst vier Publikationen, die sich speziell mit diesen Herausforderungen im Bereich der physikalisch basierten 3D-Rekonstruktion befassen.
Zunächst schlagen wir einen Variationsansatz zur Verbesserung der Auflösung einer einzelnen Tiefenkarte vor, um die gleiche Auflösung des begleitenden Farbbildes eines RGB-D-Sensors zu erreichen. Durch die Kombination von Tiefen- und Farbdaten lösen wir gleichzeitig die Probleme Tiefenkarten-super-resolution (SR) und Shape-from-Shading (SfS). Durch die genaue Modellierung der Auflösungsdiskrepanz von RGB-D-Sensoren zeigen wir, dass die Extraktion niederfrequenter geometrischer Informationen aus niedrig aufgelösten Tiefenkarten zur Disambiguierung von SfS und die Verwendung hochfrequenter photometrischer Informationen aus dem RGB-Bild zur Disambiguierung von Tiefenkarten-SR führen kann.
Zweitens bauen wir auf den Prinzipien von Photometric Stereo (PS) auf, indem wir einen effizienten Variationsansatz für unkalibriertes PS unter allgemeinen Beleuchtungsbedingungen einführen. Wir approximieren das Lambertsche Reflexionsmodell durch Kugelflächenfunktionen, was die Wiederherstellung von Geometrie, Reflektanz und Beleuchtung durch ein einziges Variationsproblem ermöglicht. Der vorgeschlagene Ansatz verwendet eine innovative Minimalflächeninitialisierung und eliminiert zusätzlich die Notwendigkeit einer nachfolgenden Normalenintegration.
Als Nächstes befassen wir uns mit dem gemeinsamen Problem der 3D-Rekonstruktion und 2D-Segmentierung von Objekten unter Verwendung von PS. Im Gegensatz zu früheren Arbeiten, die von einer vorberechneten Segmentierung des Bildes ausgehen, formulieren wir die Rekonstruktion und Segmentierung als ein gemeinsames Problem. Durch die Kombination einer Differentialformulierung von PS mit dem Chan-Vese-Modell für aktive Konturen leitet die vorgeschlagene Variationslösung simultan eine binäre Maske des Objekts und eine Tiefenkarte ab.
Abschließend präsentieren wir eine Methode zur Schätzung der Bidirektionalen Reflexionsverteilungsfunktion (BRDF) in umfangreichen Szenen von unkontrollierter Umgebung. Wir zerlegen die BRDF in diffuse und nicht-diffuse Komponenten auf und lösen jede Komponente separat. Durch den Einsatz von Monte-Carlo-Strahlenverfolgung und einer auf Strahlenverfolgung basierenden Optimierungsstrategie können wir alle Para-

meter der BRDF effizient schätzen.
Insgesamt adressieren wir entscheidende Probleme in der physikalisch basierten 3D-Rekonstruktion, wie beispielsweise Tiefenkarten-SR, SfS, unkalibriertes PS, PS und Segmentierung sowie BRDF-Schätzung. Die vorgeschlagenen Ansätze wurden anhand anspruchsvoller synthetischer und realer Daten evaluiert und zeigen im Vergleich zu bestehenden Methoden eine verbesserte Leistung, Effizienz und Robustheit, was den Weg für weitere Entwicklungen und Anwendungen in diesem Bereich ebnet.

# Acknowledgements

First and foremost, I deeply appreciate my doctoral supervisor, Daniel Cremers, for his guidance throughout my research journey. Your expertise and willingness to support my pursuit of topics aligned with my interests have significantly contributed to my academic growth. Working under your supervision has been an invaluable experience that will leave a lasting impact on my professional path. Your commitment to fostering a diverse and highly functional research group has created an outstanding work atmosphere. The presence of renowned on-site and visiting researchers has set a high standard for achieving our individual goals, benefiting everyone involved. Beyond the research setting, I have thoroughly enjoyed engaging in various out-of-office activities with you and the group. Whether it was hiking, after-work socials, barbecues, beach volleyball, laser tag, surfing, or even scuba diving, these experiences have added a memorable dimension to our professional relationship. I will never forget encountering turtles and octopuses in Hawaii and the incident when you ran out of gas. What a dive! I am genuinely thankful for your mentorship, the vibrant research environment you cultivated, and the camaraderie we shared beyond the academic realm. I will carry the lessons learned and memories forged during this time throughout my career.

I am sincerely honored to have Bastian Goldlücke as an external referee on my examination committee. Your invaluable insights and approach to addressing problems and challenges are truly inspiring. You can bridge the gap between practical vision problems and theoretical mathematical principles, offering a unique perspective that enriches the research process.

I extend my sincerest appreciation to Prof. Rüdiger Westermann, Prof. Daniel Cremers, and Prof. Bastian Goldlücke for graciously accepting the responsibility of serving as esteemed members of my examination committee.

I am thankful to Yvain Quéau, Tao Wu, Robert Maier, Mohammed Brahimi, and Matthias Vestner for proofreading my thesis and providing valuable feedback.

The chair of Computer Vision & Artificial Intelligence would not function as well without Sabine Wagner and Quirin Lohr. You make life easier for us Ph.D. students and everyone at the chair. Sabine, I can still not fill out a “Dienstreiseantrag” without your help. And Quirin, you are the Linux wizard who keeps the whole ecosystem running smoothly. I would have been lost with Ubuntu if it were not for you. Thanks again!

I am grateful for the awesome group dynamic within the chair, as all of you played a significant role in creating lasting memories during my time there. The sports team, with Marvin Eisenberger, Nikolaus Demmel, and myself, brought me to my peak physical condition. Burpees and pull-ups are not the same without you. Great after-work din-

ners, exploring unique culinary destinations with Thomas Möllenhoff, Tao Wu, and John Chiotellis. The Weisswurst breakfast table, featuring the punctuality of Robert Maier and the late arrivals of Matthias Vestner, holds cherished memories. At 4 p.m., Zorah Lähner would reliably knock on my office door, marking our customary afternoon break. The shared experiences during conference trips to Venice, Salt Lake City, Seoul, and Hawaii, alongside Yvain Quéau, Thomas Möllenhoff, Robert Maier, Rui Wang, Tao Wu, Zhenzhang Ye, Maolin Gao, Maxim Maximov, Csaba Domokos, Emanuel Laude, Florian Hofherr, and Lu Sang. I would also like to express my admiration to the entire team, including Mohammed Brahimi, Georg Kuschk, Christiane Sommer, Laura Leal-Taixé, Tim Meinhardt, Patrick Dendorfer, Virginia Estellers, Philip Häusser, Caner Hazırbaş, Mariano Jaimez Tarifa, Lingni Ma, Michael Möller, Emanuele Rodolà, Frank Schmidt, Vladimir Golkov, Vladyslav Usenko, Thomas Frerix, Nan Yang, David Schubert, Qunjie Zhou, Aysim Toker, Florian Bernard, Rudolph Triebel, Qadeer Khan, Lukas Köstler, Yuesong Shen, Lukas von Stumberg, Christian Tomani, Patrick Wenzel, and Tarun Yenamandra. Each of you has contributed significantly to my experience at the chair. Furthermore, I extend my gratitude to my former students, who have embarked on successful careers: Mohammed Brahimi and Lu Sang, who have joined Daniel's team, Alok Verma, Christoph Kick, Florian Windolf, Lisa Kaldich, and Sinem Dalkılıç.

A special thank you goes to my esteemed Ph.D. mentor and friend, Yvain Quéau, whose introduction to photometry has ignited a deep passion within me. Through your assistance, I have gained a fresh perspective on 3D reconstruction, which has captivated my interest like never before. Your outstanding research capabilities, ability to simplify complex mathematical concepts, and remarkable talent for producing eloquent and poetic papers overnight are truly astounding. Without you, I would not be where I am today. There is still much to learn from your expertise, and I am grateful for your ongoing friendship and mentorship. I also cherish the enjoyable moments shared with you outside our professional endeavors. Although many of these activities occurred late at night, lasting until the early morning hours and sometimes resulting in a regretful headache the next day, they always provided entertaining stories to share. I am sure you remember the unforgettable phrase: "I don't know where I am, but here are fu***ng lions." Furthermore, I would like to thank you for your warm hospitality during my visit to your new research group in Caen, France. Interacting with extraordinary individuals, including Jalal Fadili, was a pleasure within the group led by David Tschumperle. I want to thank Nico for making my stay memorable through engaging discussions, playing board games, and eating delicious mussels.

Throughout my doctoral venture, I had the privilege of undertaking two great research internships at Meta. I am writing to express my sincere appreciation to Thomas Whelan, my supervisor, during my initial internship. You exemplify the epitome of a backing and inspiring mentor. Your approachability, calm demeanor, and support have contributed significantly to my personal and professional growth. I am genuinely grateful for your encouragement, and the valuable insights you shared throughout our col-

laboration. These research stays provided me invaluable opportunities to engage with exceptional individuals in fruitful discussions, yielding impactful outcomes. I want to extend my sincerest respect to Jesus Briales, Simon Green, Michael Goesele, Richard Newcombe, Richard Szeliski, Daniel Andersen, Alan Oursland, Samir Aroudj, Steven Lovegrove, Eddy Ilg, and Zhao Dong.

Lastly, and of utmost importance, I express my deepest gratitude to my family. In particular, my parents, Evi and Otto, whose dedication and guidance have played an instrumental role in shaping the person I am today. Your unyielding belief in me and your immeasurable patience have propelled me to pursue and achieve my goals. Your constant presence and encouragement have given me the strength and resilience to never lose sight of my ambitions. I will forever cherish the defining phrase you instilled in me: "Wenn du etwas wirklich willst, dann schaffst du das auch!" (If you truly want something, you will achieve it).

I hold my dear brother Alex in high regard, whose profound knowledge and passion for mathematics have shaped my academic path. Your teachings have allowed me to comprehend and embrace the inherent beauty of mathematics, and I owe a great deal of my achievements to your invaluable influence.

I am grateful beyond words to my dear Ina, who has been by my side since the beginning of my doctoral voyage. You have endured the highs and lows with aid and shared in the struggles and celebrations. I sincerely admire your steadfast love, patience, and companionship during these challenging times. Your presence has been a constant source of solace and strength, for which I am profoundly thankful.

Additionally, I would like to acknowledge my children, Emilia and Dominik, who may not fully comprehend the extent of their encouragement, but whose mere presence has been a source of immeasurable inspiration. Regardless of the challenges or difficulties I encountered, a smile on your faces has effortlessly dissolved every struggle, reminding me of the greater purpose that drives me forward.

To my entire family, I am forever indebted to you for your love, encouragement, and unwavering belief in me. With immense gratitude, I conclude this phase of my journey, knowing that your steadfast support has been the cornerstone of my success.

The duration encompassing all the remarkable moments, including the interactions with diverse individuals, engaging trips, stimulating activities, pivotal breakthroughs, and formative challenges, has instilled in me a profound sense of appreciation. Now, with these experiences as a foundation, I am filled with eager anticipation for the promising prospects in my future endeavors.

— Björn

# Contents

# Part I

# Introduction and Preliminaries

# Chapter 1

# Introduction

As humans, we possess a remarkable ability to excel at the task of 3D reconstruction by effortlessly perceiving the world. We can readily estimate the source of light, recognize material when seeing a reflection, or understand an object's shape by simply looking at it. However, for machines, the task of estimating such properties from a set of 2D images is a complex challenge. In fact, in both fields computer vision and computer graphics 3D reconstruction is a fundamental problem and has a long-standing history. In recent years, we have seen a significant boost in this area due to two main phenomena. First, the availability of low-cost commodity sensors, such as RGB-D sensors and high-quality cameras in smartphones, and second the increase in computational power, which enables solving more complex and larger problems. Consequently, the demand for not only solving 3D reconstruction but also achieving high-quality reconstructions has increased. Such high-quality reconstructions are crucial for several applications, including augmented reality (AR), virtual reality (VR), or mixed reality (MR) applications in the gaming industry [38, 110], automatic visual inspection of scenes [67], computer-aided surgery tasks [50], autonomous driving [229], and smart devices [39].
The proposed thesis aims to tackle 3D reconstruction problems while achieving the aforementioned quality aspects. To this end, we will leverage relations that humans can so effortlessly process, specifically the relation between image brightness of the captured scene and its illumination, geometry, and material. These relations will then be used to robustly recover the mentioned scene assets of unprecedented detail and quality. By dedicating attention to simple yet effective formalisms, this thesis aims to provide a comprehensive framework for solving *physically-based inverse problems for high-quality 3D reconstruction.*

## 1.1 Motivation

To establish the significance of high-quality 3D reconstruction based on physical principles, this thesis will specifically motivate two aspects. The first aspect involves the critical task of accurately estimating geometry, while the subsequent will focus on the challenging problem of robustly recovering material properties.

**High-Quality Geometric Reconstruction.** While several approaches have been proposed for reconstructing geometry from either RGB data [205] or geometric data [158], few works have attempted to combine both modalities [261]. Moreover, physically-based 3D reconstruction methods are even scarcer [263]. Among the corresponding geometric reconstruction approaches, two stand out: Shape-from-Shading (SfS) [92] and photometric stereo (PS) [238]. Despite their challenges, both methods are active research areas, due to their potential for high-quality geometric estimation.
SfS is a well-studied approach for estimating shape from a single image, owing to its simplicity in problem statement and data acquisition. However, most existing methods either suffer from an inherently ambiguous problem formulation [19], or require multiple views of the scene to obtain shape estimates [263]. This limitation can cause a lack of robustness or make the capturing process more complex. Therefore, it is desirable to have a solution that uses a simple data acquisition process while still solving an unambiguous problem. One potential solution is to leverage the capabilities of modern hardware, such as RGB-D sensors. Although few approaches use a single RGB-D image pair [164], they typically fail to model sensor discrepancies, such as the difference in resolutions between the RGB and depth images. Consequently, the resulting geometry tends to have low resolution. Therefore, the low-resolution (LR) depth image is the limiting factor, as valuable high-resolution information can not be used to its full extent.
Compared to SfS, geometry reconstruction in the case of PS is less ambiguous, due to the availability of multiple images under different lighting [238]. In theory, this allows for high-detailed reconstructions based on the images alone. However, all existing works rely on a given mask, and additionally most focus on controlled laboratory settings, *e.g.*, known directional light [208] or orthographic camera projection [252]. Moreover, it is popular to reconstruct geometry via normal estimation alone [20], which may result in non-curl-free normals that do not represent a surface. These limitations make the PS approach complex in terms of data acquisition and reliant on impractical assumptions.
To address the research gaps in the domain of SfS and PS, our work in Chapters 5, 6, and 7 is aimed at utilizing RGB-D hardware, exploring depth super-resolution (SR), simultaneous masking, general lighting scenarios, realistic camera models, as well as depth reconstruction to overcome the challenge of non-integrability.

**High-Quality Material Reconstruction.** Material estimation entails the recovery of bidirectional reflectance distribution function (BRDF) parameters, enabling the rendering of photorealistic objects in images that are virtually indistinguishable from real-world photographs. This capability is of utmost importance for creating immersive scene content in AR, VR, and MR applications, as mentioned earlier.
Many existing methods operate under the assumption of a simplified Lambertian model that neglects specular highlights. While this approach may offer some advantages in terms of computational simplicity, it leads to renderings that deviate from the faithful representation of images, as reflections are an inherent and essential aspect of our world. Incorporating view-dependent material effects introduces additional complexities to the problem of material reconstruction. Even when illumination and geometry are known, recovering non-diffuse reflectance remains a challenging task, often requiring additional assumptions [74]. Consequently, in the case of non-diffuse materials, the focus is typically directed towards solitary objects within controlled environments. However, achieving a faithful rendering of large-scale scenes necessitates a substantial database of objects with accurate material properties. Creating such a database results in the reconstruction of material properties for numerous objects of varying scales which becomes a time-consuming endeavor. Conversely, in large-scale scenes, establishing controlled environments for each object presents its own significant difficulties. This inherently limits the automated estimation of a comprehensive set of BRDF parameters for each object in large-scale scenes.
Our objective in Chapter 8 is to tackle these challenges and achieve a fully automated approach for estimating BRDF parameters for every object within a large-scale, uncontrolled environment.

## 1.2 Thesis Outline

This cumulative dissertation is organized into four distinct sections, as outlined below:
**Part I** provides an in-depth introduction to the research problem that motivated this thesis, including the underlying theoretical background and methodology. The first chapter, Chapter 1, offers an introductory overview of the research topics, while Chapter 2 establishes the mathematical tools and fundamental concepts necessary for physically-based 3D reconstruction. Additionally, Chapter 3 provides a detailed literature overview of the current state-of-the-art in RGB-D depth SR, SfS, PS, masking algorithms, and BRDF parameter estimation of large-scale scenes. Finally, in Chapter 4, the main contributions of this work are presented alongside an overview of the respective peer-reviewed publications.

**Part II** contains the four original peer-reviewed publications that comprise the cumulative content of this thesis, along with their respective disclaimers. Chapter 5 presents a method for solving depth SR in combination with SfS using RGB-D data [5], resulting in high-quality depth maps of unprecedented detail. Chapter 6 introduces a novel variational approach [6] aimed at solving uncalibrated photometric stereo (UPS) under general illumination, which improves the accuracy of estimated geometry by a factor of $2$–$3\times$ by robustly solving for shape, albedo, and lighting in an alternating scheme, given an innovative minimal surface initialization. Chapter 7 eliminates the fundamental assumption on the need for a segmentation mask in PS approaches by simultaneously solving for PS and a segmentation [4] based on active contour approaches. Finally, Chapter 8 proposes an efficient method for estimating the BRDF parameters of every object in large-scale, room-sized scenes through the development of novel techniques in the area of inverse ray tracing.

**Part III** offers a comprehensive summary of the thesis in Chapter 9, which serves to synthesize the contributions made throughout this work. Additionally, in Chapter 10, we discuss general limitations and potential avenues for future research within the domains of computer vision and computer graphics, as well as opportunities for expanding upon the proposed approaches.

**Part IV** comprises the supplementary material for three of the publications included in Chapters A, B, and C. This supplementary material provides additional details and derivations, as well as experimental results, which support the conclusions presented in the respective publications.

# Chapter 2

# Theoretical Background

This chapter provides a comprehensive and detailed analysis of the theoretical background of the thesis, establishing the necessary mathematical and physical foundations for the research presented in the subsequent chapters.
We commence this chapter with an in-depth discussion of two fundamental camera models, namely the orthographic and perspective models that have been used in the works presented in Part II. The underlying assumptions, principles, benefits, and limitations of each model are explained and elaborated upon in detail.
We then delve into the hardware aspects of RGB-D cameras and the various methodologies for depth measurements. Notably, we address the challenge arising from the difference in resolution between the RGB and depth image, which naturally leads us to the issue of depth super-resolution. This section is particularly relevant to the research presented in Chapter 5.
Moving on, we discuss the rendering equation and its simplifications. As all the works presented in Part II use some form of it, we start from a general formulation and discuss emissivity, diffuse and non-diffuse bidirectional reflectance distribution functions (BRDFs), directional and natural illumination, and geometry based on normals and depth maps.
Following this, we proceed to apply the aforementioned concepts and employ the methodologies of Shape-from-Shading (SfS) and photometric stereo (PS) to invert the rendering equation. Both of these problem statements are being addressed in Chapters 5 to 7.
In the concluding section of this chapter, we examine the technique of active contour segmentation, which has proven to be an effective method for segmenting objects in images. We discuss the application of active contours, as it is used in combination with PS in the work presented in Chapter 7.

## 2.1 Camera Models

In order to perform image-based 3D reconstruction, we have to relate a pixel position with its corresponding surface point and vice versa. When the surface is opaque, there is a bijection between each pixel and a 3D point on the surface. The aim of this section is to describe two different ways to model this relation. Many of the 3D reconstruction algorithms, including the ones presented in Part II use the orthographic camera model [4, 6] or pinhole camera model [2, 5, 6]. We will discuss each model in more detail in the following sections. Other models like weak perspective projection [17, 139], spherical projection [139], enhanced unified camera [116], Kannala-Brandt camera [108], field-of-view camera [57], or the double sphere camera model [224], et cetera are out of scope of this thesis.
In general, a camera projection allows us to relate 2D image pixel positions, $\mathbf{p} \in \mathbb{R}^2$ with their corresponding 3D points, $\mathbf{x} \in \mathbb{R}^3$. Thus, we can define a projection as a mapping $\Pi : \mathbb{R}^3 \to \mathbb{R}^2$, $\mathbf{x} \mapsto \Pi(\mathbf{x}) = \mathbf{p}$. Depending on the model, the specific definition of $\Pi$ and its inverse $\Pi^{-1}$ differ.

### 2.1.1 Orthographic Camera

An *orthographic camera model* is based on orthographic projection. A particularly simple mapping, as a point $\mathbf{x} = (x, y, z)^\top \in \mathbb{R}^3$ and its pixel $\mathbf{p} = (u, v)^\top \in \mathbb{R}^2$ are related via $x = u$ and $y = v$,

$$\Pi_\mathrm{o} : \mathbb{R}^3 \to \mathbb{R}^2, \quad \mathbf{x} \mapsto \Pi_\mathrm{o}(\mathbf{x}) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} u \\ v \end{pmatrix} = \mathbf{p}. \tag{2.1}$$

The inverse of this mapping needs additional information, as the $z$ information is lost after applying $\Pi_\mathrm{o}$. Thus, we can define the inverse orthographic projection as,

$$\Pi_\mathrm{o}^{-1} : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}^3, \quad (\mathbf{p}, z) \mapsto \Pi_\mathrm{o}^{-1}(\mathbf{p}, z) = \begin{pmatrix} u \\ v \\ z \end{pmatrix} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \mathbf{x}. \tag{2.2}$$

An illustration of this camera model is depicted in Figure 2.1. This projection is a reasonable approximation of the perspective camera analyzed in the next section, if the scene's depth variation or its diameter is much smaller than the distance of the scene to the image plane [139]. Additionally, as we will see in Section 2.3.4, in order to compute a surface normal from a depth map, we need to specify a camera model. Hence, the

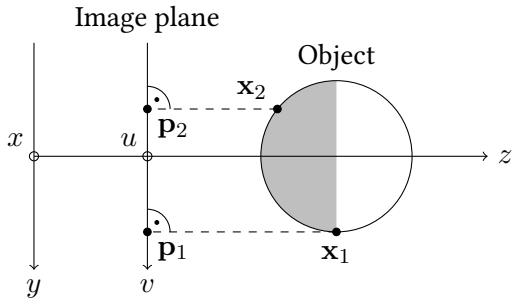



Figure 2.1: Illustration showcasing orthographic camera projection. The camera-coordinate system is represented by the $xyz$-coordinate system, while the image plane is the $uv$-plane. The gray portion of the spherical object represents the visible area captured by the camera. From the camera's perspective, point $\mathbf{x}_1$ is situated on the object's boundary, while point $\mathbf{x}_2$ lies within the object. The pixels $\mathbf{p}_1$ and $\mathbf{p}_2$ correspond to the orthographic projections of $\mathbf{x}_1$ and $\mathbf{x}_2$, respectively.

orthographic case is a popular choice for SfS and PS approaches, due to its ease of use, as shown in Chapter 6 and Chapter 7 of this thesis. In practice, no camera calibration or knowledge of the camera intrinsics is needed[1]. However, if above assumptions do not hold, or if more knowledge of the camera is available, one should resort to the more realistic perspective projection.

### 2.1.2 Perspective Camera

One of the most well-known camera models is the *perspective camera model*, also called *pinhole camera model* which is based on perspective projection. In this case, a 3D point $\mathbf{x} = (x, y, z)^\top$ is projected onto the plane $z = 1$, scaled, and shifted,

$$\Pi_\mathrm{p} : \mathbb{R}^3 \to \mathbb{R}^2, \quad \mathbf{x} \mapsto \Pi_\mathrm{p}(\mathbf{x}) = \begin{pmatrix} f_x \frac{x}{z} + c_x \\ f_y \frac{y}{z} + c_y \end{pmatrix} = \mathbf{p}. \tag{2.3}$$

The scale in $x$- and $y$-direction is called the *focal length* $f_x$ and $f_y$, whereas the shift in $x$- and $y$-direction is called the *principal point* $c_x$ and $c_y$. The perspective projection in (2.3) is an affine linear transformation in $\frac{\mathbf{x}}{z}$. With the help of homogeneous coordinates of $\mathbf{p} = (u, v)^\top \in \mathbb{R}^2$ in their projective space $\mathbb{P}^2$, $\tilde{\mathbf{p}} = (u, v, 1)^\top \in \mathbb{P}^2$, we transform (2.3) into a compact linear equation in $\frac{\mathbf{x}}{z}$,

$$\tilde{\Pi}_\mathrm{p}(\mathbf{x}) = \frac{1}{z}\mathbf{K}\mathbf{x} = \tilde{\mathbf{p}}. \tag{2.4}$$

[1]In theory, it is necessary to convert pixel units to metric units in order to determine the pixel size. However, in practice, this step is frequently ignored.

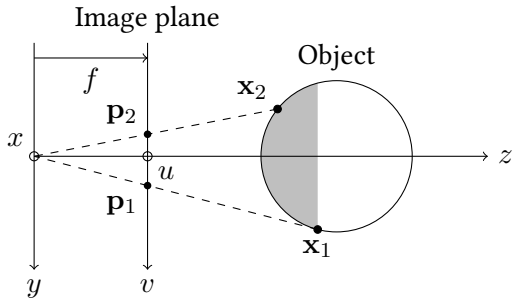



Figure 2.2: Illustration showcasing perspective camera projection. The camera-coordinate system is represented by the $xyz$-coordinate system, while the image plane at distance $f$ is the $uv$-plane[2]. The gray portion of the spherical object represents the visible area captured by the camera. From the camera's perspective, point $\mathbf{x}_1$ is situated on the object's boundary, while point $\mathbf{x}_2$ lies within the object. The pixels $\mathbf{p}_1$ and $\mathbf{p}_2$ correspond to the perspective projections of $\mathbf{x}_1$ and $\mathbf{x}_2$, respectively.

Here, $\tilde{\Pi}_\mathrm{p} : \mathbb{R}^3 \to \mathbb{P}^2 \subset \mathbb{R}^3$ indicates the mapping to the projective space $\mathbb{P}^2$. For more information on projective spaces and homogeneous coordinates, we recommend [194]. In practice, one can easily recover $\mathbf{p}$ from $\tilde{\mathbf{p}}$ via (2.1). Hence, $\Pi_\mathrm{p} = \Pi_\mathrm{o} \circ \tilde{\Pi}_\mathrm{p}$, knowing that $\mathbb{P}^2 \subset \mathbb{R}^3$. The matrix $\mathbf{K} \in \mathbb{R}^{3\times3}$ is referred to as the *intrinsic matrix*,

$$\mathbf{K} = \begin{pmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{pmatrix} \tag{2.5}$$

and its collection of parameters $\{f_x, f_y, c_x, c_y\}$ are known as the *intrinsic parameters*. Note, that we neglect a potential skew in the pixel grid, *i.e.*, all pixels are assumed to be rectangular. If we assumed a non-trivial skew coefficient, we would need to account for this in the intrinsic matrix [139].
Similar to the orthographic projection, for an inverse projection the depth information $z$ is missing. Hence, when reprojecting the pixel $\mathbf{p}$ into the 3D scene, we need to know its depth value,

$$\Pi_\mathrm{p}^{-1} : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}^3,\ (\mathbf{p}, z) \mapsto \Pi_\mathrm{p}^{-1}(\mathbf{p}, z) = z\mathbf{K}^{-1}\tilde{\mathbf{p}} = \mathbf{x}. \tag{2.6}$$

An illustration of perspective projection is depicted in Figure 2.2. As for the pinhole camera model all rays are forced to go through the optical center of the camera, it assumes a point aperture with an infinitesimal small lens. This also means that all points that lie on the same ray are projected onto the same point in the image plane [139].

[2]Figure 2.2 assumes that $f := f_x = f_y$. In cases where $f_x \neq f_y$, there is no single focal point, resulting in a phenomenon known as *astigmatism*.

The Chapters 5, 6, and 8 in this thesis rely on the pinhole camera model, under the assumption of a well-focused imaging configuration, devoid of any blur or distortion. Under these assumptions, perspective projection is a good approximation. However, in order to successfully use a perspective camera model the intrinsic parameters have to be known. These are usually either accessible via the exchangeable image file format (EXIF), have to be carefully calibrated [196], are provided by the manufacturer of the camera, or are available in third party software. For example, in the case of the RGB-D sensor Asus Xtion Pro Live, the camera intrinsics can be accessed via the third party software *OpenNi*[3]. Thus, the perspective camera model is a widely adopted assumption in the context of commodity RGB-D cameras. These sensors have had a significant impact on 3D reconstruction, which we will discuss in the subsequent section.

## 2.2 RGB-D Cameras and their Resolution Problem

During the last decade, the area of dense 3D reconstruction experienced a major boost. One significant factor was the release of the *Microsoft Kinect V1* in late 2010. It is a low-cost commodity RGB-D sensor which delivers registered image data and dense depth measurements in real-time. This led to real-time dense 3D reconstruction of objects [158]. In the same year, Stühmer *et al.* [217] developed an approach that was able to do real-time dense geometry reconstruction from gray-value images. However, the Microsoft Kinect gained popularity fast, due to its ease of use and robust, good quality depth maps. In the following, we will discuss the two most used methods for depth measurements in RGB-D sensors, namely structured light and time-of-flight (TOF). For a recent survey and in-depth discussion of depth cameras, we refer to [76].

### 2.2.1 Structured Light

Shortly after the Kinect V1 was published, other RGB-D devices were released. The most renowned and widely used ones are the *Asus Xtion Pro live*, the *PrimeSense Carmine* and the *Structure Sensor PRO*. All devices have the same underlying depth measurement approach called *structured light*. This method dates back to the 70s [209, 237] and probably the most well-known dataset based on structured light is the Middlebury dataset [201]. Many different versions to leverage structured light for range image estimation have been developed [22, 197]. The underlying objective is to recover the scene's 3D geometry, given two or more images. One way to achieve this is via triangulation, if the same 3D point is seen in multiple images. To this end, correspondences are needed, *i.e.*, po-

[3] `https://structure.io/openni`, accessed on 3rd of March, 2023 at 9.14AM.

sitions in the images that correspond to the same 3D point. One way of finding dense correspondences is to project a known structured pattern onto the scene with a projector. An observing device captures the distorted pattern and can find correspondences via computing its deformation [75]. The situation for RGB-D sensors is slightly more complicated as the projector of the structured light pattern should not visually degrade the corresponding RGB image. To alleviate this, an infrared (IR) projector and camera is used which does not interfere with the RGB camera's transmission spectrum. This setup works well in indoor scenes, where no direct sunlight is visible. On the other hand, outdoor scenarios often pose challenges for these sensors as the IR camera may struggle to detect the pattern, due to the interference caused by the intense IR radiation emitted by the sun. Other limitations of these depth measurement devices are quantization effects, partial occlusions and holes, as well as wrong measurements due to dark or very reflective surfaces. Overall the measurement error, *e.g.*, due to noise increases quadratically w.r.t. the distance, see [117]. Some of these undesirable effects can be alleviated, *e.g.*, quantization artifacts, when resorting to other techniques like time-of-flight.

### 2.2.2 Time-of-Flight

Time-of-flight (TOF)-based devices rather count to a second generation of RGB-D cameras. Popular models are the *Asus Xtion 2*, the *Microsoft Kinect V2*, and the *Microsoft Azure Kinect DK*. TOF cameras use a photonic mixer device (PMD), which was first developed in 1998 [244], hence they are also called PMD cameras. Similarly to the previous method, an emitter and a camera are needed for the TOF approach. However, instead of looking for correspondences, the time of flight is measured for an emitted light pulse to travel through the scene and back to the camera. The measured time can be used to deduce the distance to the scene. This is either done via an optical shutter approach or an intensity modulation approach [76, 120]. In the case of RGB-D cameras the TOF component also uses IR devices to not pollute the RGB images. The imaging process of TOF cameras using IR results in similar limitations as structured light sensors when capturing outdoor scenes or materials with reflective or absorbing properties. Due to the stereo vision setup partial occlusions and holes are also common in the captured data. However, in contrast to structured light systems, TOF systems exhibit fewer quantization artifacts. Nonetheless, TOF cameras can suffer from artifacts known as *flying pixels*, which can arise when a pixel captures a region with a depth discontinuity [120].

Independent of the methodology of an RGB-D sensor discussed here, they all share the same drawback in resolution. The RGB image and the depth image differ in image

(a) RGB image of $1280 \times 1024$ resolution.

(b) Depth of $640 \times 480$ resolution.

Figure 2.3: Illustration of the resolution discrepancy between RGB and depth image of an Asus Xtion Pro Live.

size. While the color image is in general of high resolution, the depth image tends to be of much smaller resolution. For instance, the Asus Xtion Pro Live provides RGB images with dimensions $1280 \times 1024$, whereas the maximum size of its depth images is limited to $640 \times 480$, as illustrated in Figure 2.3.

A list of all RGB-D devices mentioned here along with their acquisition method and resolution can be seen in Table 2.1.

| **RGB-D Model** | **Method** | **Color** | **Depth** |
|---|---|---|---|
| MS Kinect V1 | Structured Light | $640 \times 480$ | $640 \times 480$ |
| PrimeSense Carmine | Structured Light | $1280 \times 960$ | $640 \times 480$ |
| Asus Xtion Pro Live | Structured Light | $1280 \times 1024$ | $640 \times 480$ |
| Structure Sensor Pro | Structured Light | N/A (iPad dependent) | $1280 \times 960$ |
| MS Kinect V2 | TOF | $1920 \times 1080$ | $512 \times 424$ |
| Asus Xtion 2 | TOF | $2592 \times 1944$ | $640 \times 480$ |
| MS Azure Kinect DK | TOF | $4096 \times 3072$ | $1024 \times 1024$ |

Table 2.1: Comparison of method and resolution of RGB-D sensor models. Systematically, the depth resolution is lower than the RGB image resolution, or both resolutions are limited to VGA ($640 \times 480$). For further information we refer to [53, 76, 82, 123, 235].

In various applications, it is often crucial to obtain a depth map with high-resolution and fine-scale details, free of noise, quantization effects, and missing information, that is

consistent with the high-resolution RGB image counterpart. To achieve this, the problem of increasing a depth map's resolution is addressed by a technique known as depth super-resolution, which we will delve into next.

### 2.2.3 Depth Super-Resolution

In general, upsampling data in space and/or time is an active research problem. This can be done in the 1D case for audio signals [121, 130], or in a multi-dimensional case *e.g.*, temporal or spatial video super-resolution (SR) [40, 241], or light field SR [233]. Our work is more aligned with the 2D case *i.e.*, imaging data. However, we will not focus on RGB image SR [59, 225, 245, 251], but on the problem called *depth super-resolution*. The depth SR problem based on RGB-D data will be tackled in Chapter 5, thus we discuss it here in more detail.

In depth SR, the aim is to upsample a coarse depth map to a higher resolution. Mathematically, we can express this in terms of a low-resolution (LR) depth map $z_{\text{LR}} : \Omega_{\text{LR}} \to \mathbb{R}$ and an SR depth map $z_{\text{SR}} : \Omega_{\text{SR}} \to \mathbb{R}$, which maps from an LR and SR image domain $\Omega_{\text{LR}} \subset \mathbb{R}^2$ and $\Omega_{\text{SR}} \subset \mathbb{R}^2$, respectively. To express this difference in resolution, we write $\Omega_{\text{LR}} \subset \Omega_{\text{SR}}$. The theoretical idea behind the relation between $z_{\text{LR}}$ and $z_{\text{SR}}$ is that the former is a downsampled version of the latter. We can formulate this in terms of a linear downsampling operator $D : \mathbb{R}^{\Omega_{\text{SR}}} \to \mathbb{R}^{\Omega_{\text{LR}}}$,

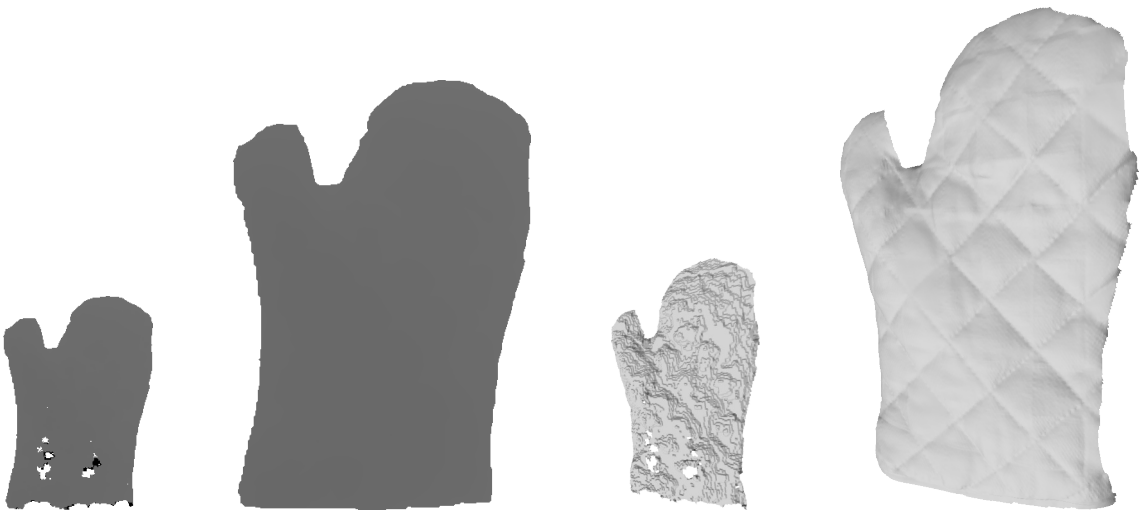
$$z_{\text{LR}} = D z_{\text{SR}} + \eta_z. \tag{2.7}$$

The operator $D$ can also incorporate other phenomena like blur or warping [63, 223], *e.g.*, in the case of having different viewpoints or if camera blur effects are being modeled. The quantity $\eta_z$ is a realization of some stochastic process, like noise, quantization or other measurement errors. While ideally, $\eta_z$ is vanishing, we saw in Section 2.2.1 and 2.2.2 that depth maps from RGB-D cameras can have non-trivial noise characteristics. In Figure 2.4, we present an illustrative pair of LR and SR depth maps, where the former exhibits visible noise contamination.

Inverting (2.7), from a given LR depth map $z_{\text{LR}}$ and a downsampling operator $D$ is an ill-posed problem due to $D$ being non-injective. Hence, additional prior terms have to be proposed in a depth SR problem,

$$\min_{z_{\text{SR}} : \Omega_{\text{SR}} \to \mathbb{R}} E_{\text{data}}(z_{\text{SR}}; z_{\text{LR}}) + \lambda E_{\text{prior}}(z_{\text{SR}}). \tag{2.8}$$

The parameter $\lambda \geq 0$ is a trade-off parameter between the data term $E_{\text{data}}$ and the prior term $E_{\text{prior}}$. Enforcing (2.7) and assuming zero-mean Gaussian noise with variance $\sigma^2$



(a) LR depth (b) SR depth (c) LR shape (d) SR shape

Figure 2.4: Comparison of LR and SR depth and shape. The presented figure showcases the oven mitt from Figure 2.3 as a depth map along with its corresponding 3D shape. While the SR depth map demonstrates improvements in terms of missing regions and resolution compared to the LR depth, the shape exhibits enhanced geometric detail, reduced noise, and mitigated quantization effects. This significant improvement in quality has been achieved through the utilization of the approach proposed by Peng *et al.* [7].

*i.e.*, $\eta_z \sim \mathcal{N}(0, \sigma^2)$, the data term in (2.8) can be written as

$$E_{\text{data}}(z_{\text{SR}}; z_{\text{LR}}) = \frac{1}{2}\|z_{\text{LR}} - Dz_{\text{SR}}\|_2^2, \tag{2.9}$$

where $\|\cdot\|_p$ for $1 \leq p < \infty$ is the $L_p$-norm, which for $p = 2$ is the standard euclidean norm, here, over the LR domain $\Omega_{\text{LR}}$.
The prior term in (2.8) can ensure smoothness of $z_{\text{SR}}$, usually via considering its gradient map $\nabla z_{\text{SR}} : \Omega_{\text{SR}} \to \mathbb{R}^2$. A well-known choice of smoothness prior is the squared $L_2$-loss or the total variation (TV) regularization,

$$E_{\text{prior}}^{L_2}(\nabla z_{\text{SR}}) = \|\nabla z_{\text{SR}}\|_2^2, \tag{2.10}$$
$$E_{\text{prior}}^{\text{TV}}(\nabla z_{\text{SR}}) = \|\nabla z_{\text{SR}}\|_1. \tag{2.11}$$

This essentially yields a smoothed out or piecewise constant solution [141], respectively, and are possibly undesired artifacts. Other popular choices of prior terms to mitigate

these effects are the Huber-loss [96, 223, 236], or a minimal surface prior [79],

$$E_{\text{prior}}^{\text{Huber}}(\nabla z_{\text{SR}}) = \begin{cases} \frac{\|\nabla z_{\text{SR}}\|_2^2}{2\varepsilon}, & \text{if } \|\nabla z_{\text{SR}}\|_2 \leq \varepsilon, \\ \|\nabla z_{\text{SR}}\|_2 - \frac{\varepsilon}{2}, & \text{if } \|\nabla z_{\text{SR}}\|_2 > \varepsilon, \end{cases} \tag{2.12}$$

$$E_{\text{prior}}^{\text{MS}}(z_{\text{SR}}) = \|A[z_{\text{SR}}]\|_1, \text{ with} \tag{2.13}$$

$$A[z_{\text{SR}}](\mathbf{p}) = \frac{z_{\text{SR}}(\mathbf{p})}{f_x f_y} \sqrt{\|\nabla_f z_{\text{SR}}(\mathbf{p})\|^2 + \left(z_{\text{SR}}(\mathbf{p}) + \langle \mathbf{p} - c, \nabla z_{\text{SR}}(\mathbf{p}) \rangle\right)^2}, \tag{2.14}$$

where $\nabla_f = \text{diag}(f_x, f_y)\nabla$ is the gradient operator scaled with a diagonal matrix of the focal lengths, and $c = (c_x, c_y)^\top$ is the principal point, see Section 2.1.2. As a regularization term, the Huber-loss function[4], as presented in (2.12), is designed to smooth small gradients while preserving strong edges. This mitigates the well-known limitation of TV regularization that favors piecewise constant solutions, which leads to staircasing artifacts. However, for perspective depth maps, a minimal surface regularization method as described in (2.13) has been proposed in [79]. Minimal surfaces are not necessarily composed of piecewise constant depth regions. A perspective minimal surface depends on the distance, as can be observed in Equation (2.14). Therefore, moving the points of the surface to the center of projection reduces surface area, which in turn mitigates the staircasing artifacts. Although not directly applied in the case of depth SR, minimal surface regularization has already been proven useful in a line of works *e.g.*, depth estimation [79], or photometry-based approaches [3, 5, 149, 188].

However, simply solving a problem in the form of (2.8) with one of the regularization terms mentioned here leads to the smoothing out of fine geometric details, as all of these regularization methods imply local smoothness of the surface. To address this, (2.8) can be supplemented with the companion RGB image in the case of RGB-D sensor data. The color image inherently contains useful and detailed geometric information that can guide the super-resolved depth map as described in a recent survey [260]. Interestingly, these image-guided depth SR approaches use only a sparse set of information from the RGB image [58, 70, 172, 173, 246], which we will see later in Chapter 3.1.1, where we discuss each method's advantages, disadvantages, and open problems.

One promising alley of research involves the development of physically-based models that establish dense relationships between RGB images and a scene's assets, such as reflectance, illumination, and geometry through the rendering equation. These approaches have exhibited promising results in guiding the depth SR problem, as evidenced by previous works including [3, 5, 7, 8, 137]. Our work in Chapter 5 also employs this methodology, combining depth SR with SfS. However, before discussing this further, it is neces-

[4]Named after the Swiss statistician Peter Jost Huber (1934).

sary to first define and explain SfS and its relationship to the rendering equation. In the following sections, we will delve into the rendering equation, its assets, simplifications, SfS as well as PS.

## 2.3 Physically-Based Rendering

This section aims to cover a range of concepts that will be utilized in Part II, all of which fall under the umbrella of physically-based rendering. This field is focused on establishing the relationships between a scene's geometry, lighting, and material via light transport. We begin with introducing the well-known rendering equation, followed by a brief overview of emissivity. Afterwards, we introduce the concept of the bidirectional reflectance distribution function (BRDF) as well as various illumination scenarios and finally discuss geometric aspects.

The underlying principle of physically-based rendering is the celebrated *rendering equation* which was simultaneously proposed in $1986$ by James T. Kajiya [106] and David S. Immel *et al.* as the result of his Master's thesis [99, 100]. This equation models light transport at a point $\mathbf{x} \in \mathbb{R}^3$ in direction $\omega_\mathrm{o} \in \mathbb{S}^2$, where $\mathbb{S}^2 = \{\mathbf{x} \in \mathbb{R}^3 \mid \|\mathbf{x}\|_2 = 1\}$ is the unit sphere, as

$$L_\mathrm{o}(\mathbf{x}, \omega_\mathrm{o}) = L_\mathrm{e}(\mathbf{x}, \omega_\mathrm{o}) + \int_{\mathbb{S}^2} f_\mathrm{BRDF}(\mathbf{x}, \omega_\mathrm{i}, \omega_\mathrm{o}) L_\mathrm{i}(\mathbf{x}, \omega_\mathrm{i}) \max(0, \langle \mathbf{n}(\mathbf{x}), \omega_\mathrm{i} \rangle) \, \mathrm{d}\omega_\mathrm{i}. \tag{2.15}$$

The quantity $L : \mathbb{R}^3 \times \mathbb{S}^2 \to \mathbb{R}_0^+$ at a point $\mathbf{x} \in \mathbb{R}^3$ in direction $\omega_\mathrm{o} \in \mathbb{S}^2$ models radiance, which is defined as the radiant flux[5] per surface area and per solid angle. We indicate the outgoing, emitting, and incoming radiance with $L_\mathrm{o}$, $L_\mathrm{e}$, and $L_\mathrm{i}$, respectively. Emitting radiance is described in Section 2.3.1, while incoming radiance is further explained in Section 2.3.3. The integral's integrand consists of the bidirectional reflectance distribution function $f_\mathrm{BRDF}$, the incoming radiance $L_\mathrm{i}$, and the clamped dot product between the surface normal $\mathbf{n}(\mathbf{x}) \in \mathbb{S}^2$ at $\mathbf{x}$ and the incoming direction $\omega_\mathrm{i}$, $\max(0, \langle \mathbf{n}(\mathbf{x}), \omega_\mathrm{i} \rangle)$. The BRDF models how light is reflected at an opaque surface and we will discuss this in more detail in Section 2.3.2. The clamped dot product is reliant on the scene's geometry (Section 2.3.4), where the $\max$ operator encodes *self shadows*, which arise when the light direction $\omega_\mathrm{i}$ originates from behind the surface. This occurs when the dot product is negative, signifying that the angle between $\mathbf{n}$ and $\omega_\mathrm{i}$ is greater than $90°$. Furthermore, it models the spread of incident illumination over the surface at a given angle. We can get rid of the clamping operation *i.e.*, the $\max$ operator, if we restrict the integration domain to the upper hemisphere oriented along $\mathbf{n}$, $\mathbb{H}_\mathbf{n} = \{\omega \in \mathbb{S}^2 \mid \langle \mathbf{n}, \omega \rangle \geq 0\}$. The rendering

[5] Radiant flux is the energy of photons per second [15].

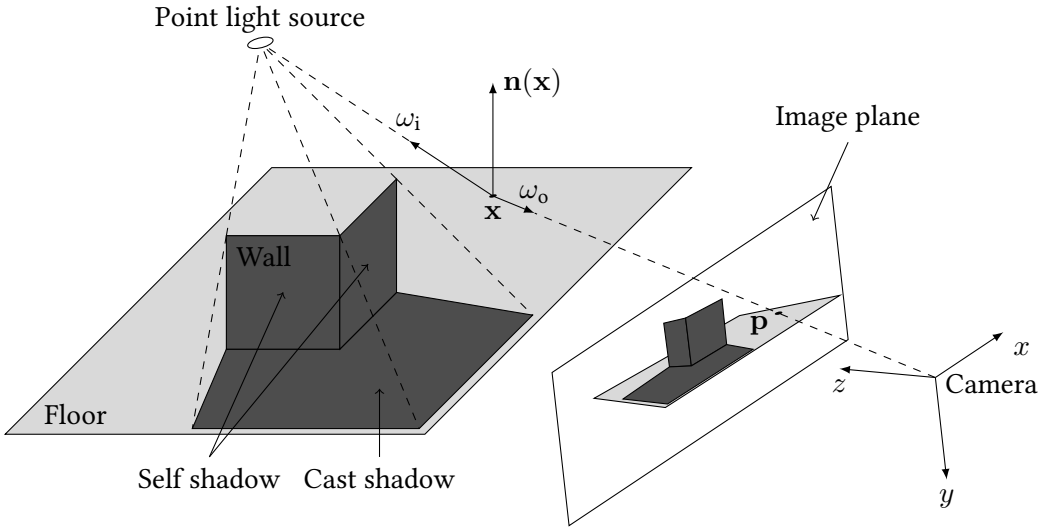



Figure 2.5: Illustration demonstrating cast shadows versus self shadows *and* the rendering equation. An opaque wall is illuminated by a point light source, resulting in shadows cast both on itself and on the floor. The facets of the wall facing away from the light source produce self shadows, while the shaded regions on the floor represent cast shadows. Cast shadows occur when incoming radiance in these areas diminishes due to obstruction by the wall, blocking the light rays. Additionally, the exemplification of the relationship between a pixel and the underlying scene properties illustrates the rendering equation. A ray originates from the center of a perspective camera, passes through the pixel $\mathbf{p}$, and intersects the scene at $\mathbf{x}$ with its surface normal $\mathbf{n}(\mathbf{x})$. At $\mathbf{x}$, in the outgoing direction $\omega_\mathrm{o}$, the outgoing radiance $L_\mathrm{o}(\mathbf{x}, \omega_\mathrm{o})$ results in the image intensity $I(\mathbf{p})$ at $\mathbf{p}$, as described in Equation (2.16). Simultaneously, the incoming radiance $L_\mathrm{i}(\mathbf{x}, \omega_\mathrm{i})$ from the point light source, arriving from the incoming direction $\omega_\mathrm{i}$, illuminates the point $\mathbf{x}$. The reflectance of the floor $f_\mathrm{BRDF}(\mathbf{x}, \omega_\mathrm{i}, \omega_\mathrm{o})$ is contingent upon its material properties, such as carpet, parquet, tiles, and so forth.

model presented in Equation (2.15) encompasses another type of shadow known as *cast shadows*. Such local lighting phenomena, which include inter-reflections as well, are brought about by the dependence of $\mathbf{x}$ in $L_\mathrm{i}$. A visual comparison of cast versus self shadows is depicted in Figure 2.5.

The outgoing radiance $L_\mathrm{o}(\mathbf{x}, \omega_\mathrm{o})$ can be set in relation with the image irradiance[6] $I$ at pixel $\mathbf{p}$, $I(\mathbf{p})$ [93]. Given a camera model (Section 2.1), we can relate a pixel position $\mathbf{p}$ with its corresponding 3D point $\mathbf{x}$, $\mathbf{p} = \Pi(\mathbf{x})$. Leveraging this, we can write

$$I(\Pi(\mathbf{x})) = L_\mathrm{o}(\mathbf{x}, \omega_\mathrm{o}), \tag{2.16}$$

where the direction $\omega_\mathrm{o}$ is the normalized vector pointing from the point $\mathbf{x}$ towards its

[6]Irradiance is the radiant flux per surface area.

corresponding pixel $\mathbf{p}$. Equation (2.16) is in fact a proportionality relationship, *i.e.*, image irradiance is directly proportional to scene radiance [93]. The proportionality coefficient is determined by the camera's characteristics, but since it is constant for all pixels, it will be disregarded in the subsequent analysis. An illustration of the rendering equation and its relation to images is visualized in Figure 2.5.

We would like to point out that we have not explicitly modeled the camera pipeline or any similar components. If our intention was to capture a more comprehensive relationship than (2.16), then the camera pipeline could potentially involve the following steps, though this list is not exhaustive: lens effects like vignetting, the integration of the transmission spectrum of the camera over the wavelengths, some preprocessing (black light substraction, normalization), white balancing, demosaicing, some postprocessing (noise reduction, sharpening), color transformation, color rendering (tone mapping, color manipulation), displaying, compressing, and storing. We refer to [55, 93, 109] for more details on the camera pipeline and its effects on the final image. Works in the area of using properly calibrated data have been developed and it was shown that this can improve certain tasks [25, 64, 152]. However, for the sake of this thesis we assume (2.16) holds true.

The rendering equation can be extended to encompass time and/or wavelength dependencies in general. Time dependency is particularly relevant in dynamic scenes where a single frame can be computed by setting a fixed time, and motion blur can be accounted for by integrating over a specific time interval. In the same vein, with the radiance mapping only to $\mathbb{R}_0^+$ sampling or integration over various wavelengths can result in a polychromatic sample, such as an RGB color triplet. However, investigating these time and wavelength dependencies to more extent than described is outside the scope of this thesis. Concerning RGB data, we presume that there is a channel-wise relationship between an image and the rendering equation. In other words, we evaluate the rendering equation separately for each color channel. Depending on the radiance and the BRDF, different scenarios may arise: a grayscale case, where both radiance and BRDF are achromatic, and three RGB cases, where either radiance or BRDF is RGB while the other is achromatic, or both are RGB. Unless specified otherwise, we will present most equations in the grayscale case, and note that the RGB case can usually be computed in a straightforward manner as mentioned above.

The rendering equation is a recursive equation, as the incoming radiance $L_\mathrm{i}(\mathbf{x}, \omega_\mathrm{i})$ at a point $\mathbf{x}$ in direction $\omega_\mathrm{i}$ is equal to the outgoing radiance $L_\mathrm{o}(\tilde{\mathbf{x}}, -\omega_\mathrm{i})$ at a *hit point* $\tilde{\mathbf{x}}$ of the scene in negative direction $\omega_\mathrm{i}$, $L_\mathrm{i}(\mathbf{x}, \omega_\mathrm{i}) = L_\mathrm{o}(\tilde{\mathbf{x}}, -\omega_\mathrm{i})$. The hit point can be understood as the first intersection point of a ray defined by the tuple $(\mathbf{x}, \omega_\mathrm{i})$ with the surface. The recursive property of (2.15) is a curse and blessing at the same time. While it allows to have a single equation describing the radiance of almost arbitrary scenes, it requires

evaluating a complex integral whose integrand depends on the evaluation of the radiance at a different point in different directions[7]. Making this evaluation tractable is still an active research topic and ideas consider Monte Carlo methods [106] to efficiently approximate the integral in (2.15), graphics processing unit (GPU) programming [175] to speed up computation utilizing parallelism, bounding volume hierarchy [83] to accelerate ray hit point computation, and many more. With state-of-the-art hardware it is already possible to evaluate (2.15) in real-time *e.g.*, in consumer video games[8]. We refer the interested reader to [15, 69, 159, 176, 177] for in detail discussions on the whole topic of the rendering equation and its efficient implementation.

In the upcoming sections of this chapter, we will comprehensively discuss each component of the rendering equation (2.15). We will initiate with the emissivity and its assumptions in this thesis, followed by an in-depth discussion of BRDFs and their diffuse and non-diffuse reflectance properties. Furthermore, we will explore two lighting scenarios - directional and natural lighting. Subsequently, we will shift our focus towards the geometry of the scene, and discuss normal and depth maps and their interrelation. This comprehensive understanding of the forward model will serve as a foundation for its subsequent inversion using the techniques of SfS and PS.

### 2.3.1 Emissivity

The emissivity term, denoted as $L_{\mathrm{e}}(\mathbf{x}, \omega_{\mathrm{o}})$ in the rendering equation (2.15), describes the amount of radiance emitted by a surface point $\mathbf{x}$ in a given direction $\omega_{\mathrm{o}}$. Examples of emissive objects include light bulbs, flashlights, candles, the sun, et cetera. However, in this thesis, we focus on recovering non-emissive objects,

$$L_{\mathrm{e}}(\mathbf{x}, \omega_{\mathrm{o}}) \equiv 0. \tag{2.17}$$

The assumption of non-emissivity is a commonly adopted practice in photorealistic $3$D reconstruction [2, 30, 138, 255, 256]. As a result of this assumption, the rendering equation (2.15) solely consists of the integral term, which can be computed based on the BRDF (reflectance), the incoming radiance (illumination), and the underlying shape (geometry). In the forthcoming sections, we will discuss these three quantities in greater detail, commencing with reflectance.

[7] The recursive tracing of rays is called path tracing [106]

[8] *Shadow of the Tomb Raider* (2018) was the first video game with ray tracing capabilities to render realistic shadows. *Battlefield V* (2018) was the first game that used ray tracing to render reflections. *Metro Exodus* (2019) was the first game to render global illumination. Finally, *Quake II RTX* (2019) was the first game to deploy ray tracing for shadows, reflections *and* global illumination

### 2.3.2 Reflectance

In this section we discuss the famous *bidirectional reflectance distribution function (BRDF)*, which was first proposed by Fred Nicodemus in 1965 [160]. Other commonly used names for the BRDF are *reflectance* or *material*. Intuitively, the BRDF describes how much light is reflected due to the material properties of an opaque object[9], *e.g.*, diffuse materials reflect light uniformly in all directions, while non-diffuse materials reflect light predominantly in one direction. Thus, the BRDF is a function of two directions, the incoming (light) and the outgoing (viewer) direction. A physically realistic BRDF, $f_{\text{BRDF}} : \mathbb{S}^2 \times \mathbb{S}^2 \to \mathbb{R}_0^+$ has to fulfill two properties:

- Helmholtz reciprocity:

$$f_{\text{BRDF}}(\omega_\text{i}, \omega_\text{o}) = f_{\text{BRDF}}(\omega_\text{o}, \omega_\text{i}) \tag{2.18}$$

- Energy conservation:

$$\forall \omega_\text{o} : \int_{\mathbb{H}_\mathbf{n}} f_{\text{BRDF}}(\omega_\text{i}, \omega_\text{o}) \langle \mathbf{n}, \omega_\text{i} \rangle \, \mathrm{d}\omega_\text{i} \leq 1 \tag{2.19}$$

It is worth to mentioning that the BRDF in Equation (2.15) not only depends on the incoming and outgoing light directions but also varies with the surface point $\mathbf{x}$. Such a BRDF is commonly referred to as a spatially varying BRDF (SVBRDF) and is mathematically defined as a function $f_{\text{BRDF}} : \mathbb{R}^3 \times \mathbb{S}^2 \times \mathbb{S}^2 \to \mathbb{R}_0^+$. Although the BRDF and SVBRDF differ in terms of their dependency on the 3D point $\mathbf{x}$, we may use these terms interchangeably since the inclusion or exclusion of $\mathbf{x}$ readily distinguishes one from the other. To ensure physical realism, it is imperative that both properties of a BRDF hold true for all points $\mathbf{x}$ of an SVBRDF. While it is important to bear in mind the aforementioned attributes, it should be noted that some BRDF models may not satisfy these properties. In certain cases, the energy conservation requirement of the BRDF may be relaxed to achieve faster computation [29, 178] or to enable the representation of a diverse range of materials using a single BRDF model with a limited number of parameters [38].
A widely accepted convention, which we shall adhere to, is that according to Shafer [206] the reflectance can be dichromatically represented as the sum of a diffuse (Lambertian) BRDF, $f^\text{d}_{\text{BRDF}}$, and a non-diffuse BRDF, $f^\text{s}_{\text{BRDF}}$,

$$f_{\text{BRDF}}(\mathbf{x}, \omega_\text{i}, \omega_\text{o}) = f^\text{d}_{\text{BRDF}}(\mathbf{x}) + f^\text{s}_{\text{BRDF}}(\mathbf{x}, \omega_\text{i}, \omega_\text{o}). \tag{2.20}$$

[9] In the case of non-opaque or translucent materials, incorporating a bidirectional transmittance distribution function (BTDF) is necessary. However, this topic is not within the scope of this thesis.

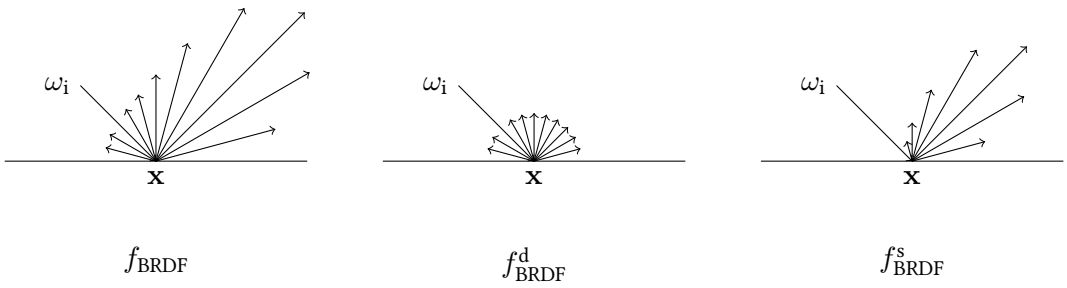



Figure 2.6: Illustration of the BRDF. The complete BRDF comprises both a diffuse and a non-diffuse component, as depicted in Equation (2.20). In this visualization, we observe a cross-section of the BRDF at a specific point $\mathbf{x}$, where the length of the outgoing directions $\omega_o$ represents the magnitude of the BRDFs for a given incoming direction $\omega_i$. While the diffuse lobe uniformly scatters light across the upper hemisphere, the specular lobe predominantly directs light in a specific direction.

The non-diffuse component of the BRDF is frequently referred to as the specular BRDF. An illustration of a purely diffuse and a purely specular BRDF, as well as their composition as described in (2.20), is shown in Figure 2.6. The specular BRDF is notable for its dependence on the light and viewing direction, whereas the Lambertian diffuse BRDF is invariant to these factors. The ensuing two sections will delve into the implications of this observation. Cook *et al.* [51] propose a convex combination of $f^{\text{d}}_{\text{BRDF}}$ and $f^{\text{s}}_{\text{BRDF}}$ to satisfy the energy conservation constraint (2.19). Therefore, depending on the specific choices of $f^{\text{d}}_{\text{BRDF}}$ and $f^{\text{s}}_{\text{BRDF}}$, the dichromatic reflectance assumption in (2.20) may not necessarily hold as energy conserving, even if $f^{\text{d}}_{\text{BRDF}}$ and $f^{\text{s}}_{\text{BRDF}}$ are each energy conserving on their own.

Over the years, a variety of BRDFs have been developed, with the Lambertian BRDF being the most well-known for $f^{\text{d}}_{\text{BRDF}}$, and the microfacet BRDF being the most well-known for $f^{\text{s}}_{\text{BRDF}}$. The Torrance-Sparrow [221] and Cook-Torrance [51] are widely used microfacet models that are known for their physical realism. In contrast, other models such as Ward [234], Phong [178], and Blinn-Phong [29], while popular due to their simplicity, are neither microfacet models nor physically realistic. Additionally, there are some microfacet BRDFs that are considered unrealistic, such as the Disney BRDF [38] which violates the energy conservation constraint in Equation (2.19).

All the core publications presented in this dissertation adhere to the dichromatic assumption (2.20). In Chapters 5, 6, and 7, we further simplify the dichromatic assumption by adopting a purely Lambertian BRDF with $f^{\text{s}}_{\text{BRDF}} \equiv 0$. However, in Chapter 8, we incorporate a non-zero specular BRDF into the model. In the forthcoming two sections, we shall concentrate on the Lambertian BRDF as a simple model for $f^{\text{d}}_{\text{BRDF}}$, and a microfacet-based model inspired by the Disney BRDF [38] for $f^{\text{s}}_{\text{BRDF}}$.

#### 2.3.2.1 Lambertian BRDF

As stated previously, throughout the majority of this thesis, it is assumed that the BRDF $f_{\text{BRDF}}$ remains invariant w.r.t. both the incident lighting and outgoing viewing direction,

$$f^{\text{s}}_{\text{BRDF}}(\mathbf{x}, \omega_{\text{i}}, \omega_{\text{o}}) \equiv 0 \tag{2.21}$$

$$\implies f_{\text{BRDF}}(\mathbf{x}, \omega_{\text{i}}, \omega_{\text{o}}) = f^{\text{d}}_{\text{BRDF}}(\mathbf{x}), \tag{2.22}$$

implying that the material's ability to reflect light is equal in all directions. It is worth mentioning that there exist alternative diffuse models, such as Oren-Nayar [165] that incorporate a dependence on $\omega_{\text{i}}$ and $\omega_{\text{o}}$. However, in our approach, we adhere to the aforementioned assumption. In this case, (2.22) is commonly referred to as *Lambertian* reflectance. This relation causes that view-dependent phenomena, such as specular highlights, are not taken into consideration in the model. Consequently, any reflections that may be present in actual images can be treated as anomalies, which can be addressed using techniques such as those described in [190].

To obtain the outgoing radiance from a perfectly diffuse surface under uniform unit radiance lighting, we can substitute non-emissivity (2.17) and the diffuse assumption (2.22) into the rendering equation (2.15), while assuming unit incoming radiance, $L_{\text{i}}(\mathbf{x}, \omega_{\text{i}}) \equiv 1$. Incorporating the fundamental principle of energy conservation, as expressed in (2.19), leads to the definition of the *albedo*, $\rho : \mathbb{R}^3 \to [0, 1]$, which describes the response of a perfectly diffuse surface illuminated uniformly by a light source with unit radiance [15]. We can express the diffuse BRDF, $f^{\text{d}}_{\text{BRDF}}$, mathematically in terms of the albedo $\rho$ using

$$\rho(\mathbf{x}) = f^{\text{d}}_{\text{BRDF}}(\mathbf{x}) \int_{\mathbb{S}^2} \max(0, \langle \mathbf{n}(\mathbf{x}), \omega_{\text{i}} \rangle) \, \mathrm{d}\omega_{\text{i}} \underset{(2.19)}{\leq} 1 \tag{2.23}$$

$$= f^{\text{d}}_{\text{BRDF}}(\mathbf{x}) \int_{\mathbb{H}_{\mathbf{n}}} \langle \mathbf{n}(\mathbf{x}), \omega_{\text{i}} \rangle \, \mathrm{d}\omega_{\text{i}} \tag{2.24}$$

$$= f^{\text{d}}_{\text{BRDF}}(\mathbf{x}) \pi \tag{2.25}$$

$$\iff f^{\text{d}}_{\text{BRDF}}(\mathbf{x}) = \frac{\rho(\mathbf{x})}{\pi}, \tag{2.26}$$

where the integral of the dot product over the hemisphere in (2.24) evaluates to $\pi$. In the context of a purely diffuse environment, the normalization factor $\pi$ can be absorbed into the proportionality coefficient derived from the relationship in Equation (2.16). Therefore, moving forward, we will disregard this factor.

As demonstrated earlier, in the case of Lambertian surfaces, the BRDF remains invariant to the integration variable of the rendering equation. Consequently, the BRDF can be extracted from the integral, leading to a simplification of the rendering equation's

integrand. We define the remaining integral as the *shading*, $S : \mathbb{R}^3 \to \mathbb{R}_0^+$ of the scene,

$$S(\mathbf{x}) = \int_{\mathbb{S}^2} L_\mathrm{i}(\mathbf{x}, \omega_\mathrm{i}) \max(0, \langle \mathbf{n}(\mathbf{x}), \omega_\mathrm{i} \rangle) \, \mathrm{d}\omega_\mathrm{i}, \tag{2.27}$$

which only incorporates light and geometric information. Using the three statements above, we can plug the non-emissivity (2.17), the diffuse reflectance (2.22), and the shading (2.27) into the rendering equation (2.15) and write it as

$$L_\mathrm{o}(\mathbf{x}) = f_{\mathrm{BRDF}}^{\mathrm{d}}(\mathbf{x}) S(\mathbf{x}). \tag{2.28}$$

The case of diffuse reflectance is important for all chapters of Part II of this manuscript. In the case of non-diffuse reflectance we resort to a microfacet BRDF strongly related to a simplification of the Disney BRDF [38], which we will discuss next.

#### 2.3.2.2 Specular BRDF

One popular approach for modeling specular reflections related to materials is through the use of the *microfacet model*, which was introduced in seminal works by Torrance and Sparrow [51] and Cook and Torrance [221]. This BRDF model describes surfaces that are not perfectly smooth, but instead composed of a multitude of randomly oriented planar surface fragments, known as microfacets. The microfacets whose orientation is halfway between the light direction and the viewing direction are responsible for the visible light reflection. However, not all of these properly oriented microfacets contribute to reflected light due to masking and shadowing effects that are accounted for in the microfacet BRDF model. That being said, a microfacet BRDF typically takes into account the half vector $\mathbf{h} = \frac{\omega_\mathrm{i} + \omega_\mathrm{o}}{\|\omega_\mathrm{i} + \omega_\mathrm{o}\|_2}$ and the surface normal $\mathbf{n}$ in the following manner,

$$f_{\mathrm{BRDF}}^{\mathrm{s}}(\omega_\mathrm{i}, \omega_\mathrm{o}) = \frac{D(\mathbf{h}) F(\mathbf{h}, \omega_\mathrm{o}) G(\omega_\mathrm{i}, \omega_\mathrm{o})}{4 \langle \mathbf{n}, \omega_\mathrm{i} \rangle \langle \mathbf{n}, \omega_\mathrm{o} \rangle}. \tag{2.29}$$

The three functions $D, F$, and $G$ are called *normal distribution function*, *Fresnel*, and *geometric shadowing*, respectively. $D$ describes the distribution of microfacets for the surface, while $G$ describes the shadowing from the microfacets, and $F$ describes the amount of light that reflects from a mirror surface.
BRDFs can be anisotropic or isotropic, which are distinguished by their invariance to rotations around the surface normal. Anisotropic BRDFs, which are not invariant to such transformations, can model non-circular specular lobes around a fixed normal direction, such as brushed metal, while isotropic BRDFs can only model circular lobes. In this work, we will focus solely on isotropic reflections as anisotropic BRDFs are beyond the scope

of our research. Furthermore, we will only consider dielectric/non-metallic surfaces as we do not aim to model metallic materials.
In Chapter 8, we utilize the microfacet model (2.29) with a specific selection of the normal distribution function $D$, the Fresnel function $F$, and the geometric shadowing function $G$.
For the surface normal distribution function $D$, we use the Trowbridge-Reitz (GGX) distribution, which additionally depends on the surface's roughness $\hat{\varphi}$ [222, 230][10],

$$D(\mathbf{h};\hat{\varphi}) = \frac{\hat{\varphi}^2}{\pi\big(\langle \mathbf{n},\mathbf{h}\rangle^2(\hat{\varphi}^2-1)+1\big)^2}. \tag{2.30}$$

The roughness parameter characterizes the surface microstructure. Surfaces with lower roughness exhibit more facets that align with the incoming light, resulting in pronounced specular reflections. In contrast, surfaces with higher roughness have fewer facets aligned with the incoming light, which leads to scattered reflections that appear blurry and diffuse.
For the Fresnel term $F$ we adopt Schlick's approximation [203],

$$F\Big(\mathbf{h},\omega_\mathrm{o};\tilde{\psi}\Big) = \tilde{\psi} + \Big(1-\tilde{\psi}\Big)(1-\langle \mathbf{h},\omega_\mathrm{o}\rangle)^5, \tag{2.31}$$

with the *specular albedo* $\tilde{\psi}$, which is related to the index of refraction (IOR).
The geometric shadow term $G$ is a function of the roughness $\tilde{\varphi}$, similar to $D$. It is common to employ the method of Smith [210] to compute $G$, which breaks it into light and view direction components and computes a pointwise product using the same function,

$$G(\omega_\mathrm{i},\omega_\mathrm{o};\tilde{\varphi}) = G_1(\omega_\mathrm{i};\tilde{\varphi})G_1(\omega_\mathrm{o};\tilde{\varphi}) \tag{2.32}$$

where $G_1$ is also based on the Trowbridge-Reitz (GGX) distribution [222, 230],

$$G_1(\omega;\tilde{\varphi}) = \frac{2\langle \mathbf{n},\omega\rangle}{\langle \mathbf{n},\omega\rangle + \sqrt{\tilde{\varphi}^2+(1-\tilde{\varphi}^2)\langle \mathbf{n},\omega\rangle^2}}. \tag{2.33}$$

It is worth noting that with this BRDF model, the denominator in (2.29) is cancelled due to the pointwise product in (2.32) and the numerator in (2.33).
Conventionally, a single roughness value for both $\hat{\varphi}$ and $\tilde{\varphi}$ is used, which is usually de-

[10]This particular form of microfacet distribution was originally developed by Trowbridge and Reitz in 1975 [222] and later reinvented by Walter *et al.* [230] in 2007, who coined the term GGX. There is an interesting post on Matt Pharr's blog about this: `https://pharr.org/matt/blog/2022/05/06/trowbridge-reitz`, accessed on 13th of April, 2023 at 3.23PM.

noted as $\varphi$. However, we depart from this convention and distinguish between the two, as we will explain in the next paragraph where we establish a connection to the Disney BRDF [38]. The Trowbridge-Reitz (GGX) distribution [222, 230] and Schlick's approximation [203] have been employed in our approach due to their successful application in the more sophisticated Disney BRDF [38]. This model is known for its capability to effectively represent a wide variety of distinct materials..

**Connection to the Disney BRDF.** We will now explore that the utilized non-diffuse BRDF mentioned above is a specific instance of the well-known Disney BRDF [38]. For the full model and thorough reasoning of the Disney BRDF, we refer the reader to [38] and the official implementation available on GitHub[11]. The complete Disney BRDF encompasses eleven parameters that account for various phenomena. The parameters are called baseColor, subsurface, metallic, specular, specularTint, roughness, anisotropic, sheen, sheenTint, clearcoat, and clearcoatGloss. As our concern is solely with the isotropic, non-diffuse, and non-metallic component of the Disney BRDF, we assign zero values to all parameters that do not contribute to this segment. The remaining terms consist of two specular lobes, the primary and the secondary, with the secondary lobe representing a thin, translucent layer, which we omit. These simplifications result in a final set of parameters, the roughness and the specular parameter.
Upon closer inspection of this simplified Disney BRDF, it is evident that the surface normal distribution function $D$, geometric shadowing $G$, and Fresnel term $F$ align with the aforementioned models, specifically equations (2.30) – (2.33), which employ the Trowbridge-Reitz (GGX) and Schlick-Fresnel models. However, a slight distinction can be observed in the approach of the Disney BRDF. It specifically involves a couple of reparameterizations of the roughness value $\varphi$, and the specular parameter $\psi$. The Disney BRDF employs these modifications to achieve improved numerical stability and a more perceptually linear change in the roughness[12],

$$\tilde{\varphi}(\varphi) = \hat{\varphi}(\varphi) = \max\big(0.001, \varphi^2\big). \tag{2.34}$$

Moreover, the specular albedo is scaled to cover most common materials,

$$\tilde{\psi}(\psi) = 0.08\psi. \tag{2.35}$$

[11] `https://github.com/wdas/brdf/blob/main/src/brdfs/disney.brdf`, accessed on April 14th, 2023 at 5:06PM.

[12] In Section 8.3.1, $\varphi^2$ from Equation (2.34) is missing. The requirement to maintain the accepted version of the publication prevents its correction.

In the initial version of the Disney BRDF manuscript, an additional reparameterization was suggested, which is employed in Chapter 8. Although $\hat{\varphi}$ is defined according to (2.34), a different expression was used for $\tilde{\varphi}$,

$$\tilde{\varphi}(\varphi) = \left(\frac{\varphi}{2} + \frac{1}{2}\right)^2. \tag{2.36}$$

This proposed third reparameterization of the roughness in the initial manuscript of the Disney BRDF was later retracted following a subsequent study [89]. Additional information on this topic can be found in [89], the addendum included in [38], and in the commit history of Disney's official GitHub repository[13]. The presented non-diffuse BRDF model has been effectively utilized in Chapter 8, wherein the reflectance characteristics of each object in a scene of significant scale have been successfully estimated via inverting the rendering equation as depicted in (2.15).

In the upcoming section on illumination, we adopt a purely diffuse BRDF model, which enables us to express the rendering equation in terms of the albedo and the shading integral, as demonstrated in (2.28). With this adoption, the shading integral $S$ can be significantly simplified under specific illumination conditions. These simplifications have been deployed in the remaining Chapters 5 – 7 of this dissertation. To this end, we will explore two frequently used lighting setups: directional lighting and natural lighting.

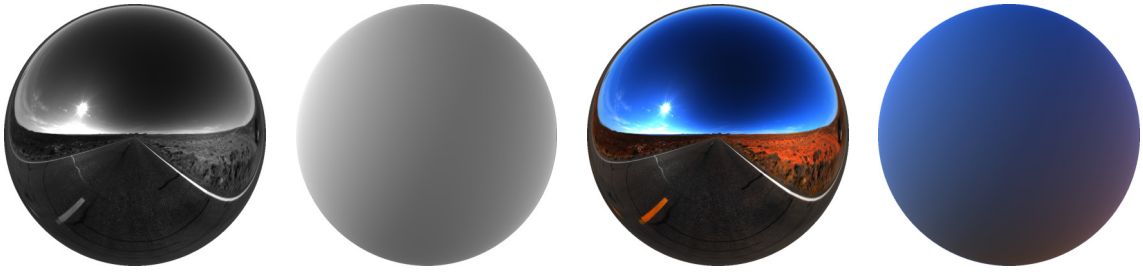
### 2.3.3 Illumination

This section addresses the incoming radiance of the rendering equation, which can be referred to interchangeably as illumination, light, or lighting. We introduce a set of assumptions regarding the incoming radiance, and demonstrate how these can be leveraged to streamline the shading process, obviating the need for explicit integration computation. It should be noted that, throughout this section, we are specifically considering diffuse reflectance (as outlined in Section 2.3.2.1). Additionally, we posit that the incoming radiance, denoted by $L_\mathrm{i}$, is positionally independent, a condition commonly referred to as the *distant light assumption*,

$$L_\mathrm{i}(\omega_\mathrm{i}) = L_\mathrm{i}(\mathbf{x}, \omega_\mathrm{i}). \tag{2.37}$$

As a result of this assumption, incoming radiance from any particular direction is identical across every point in the scene. The primary consequences of this simplification

[13] `https://github.com/wdas/brdf/commit/9aee63621cbec6891b20d6485d7b8f4549f3db1b`, accessed on 14th of April, 2023 at 5.32PM.



(a) Grayscale radiance $L_\mathrm{i}$ showing a road in Monument Valley[14].

(b) Shading $S$ of a white hemisphere using the radiance map shown in (a).

(c) RGB radiance $L_\mathrm{i}$ showing a road in Monument Valley[14] .

(d) Shading $S$ of a white hemisphere using the radiance map shown in (c).

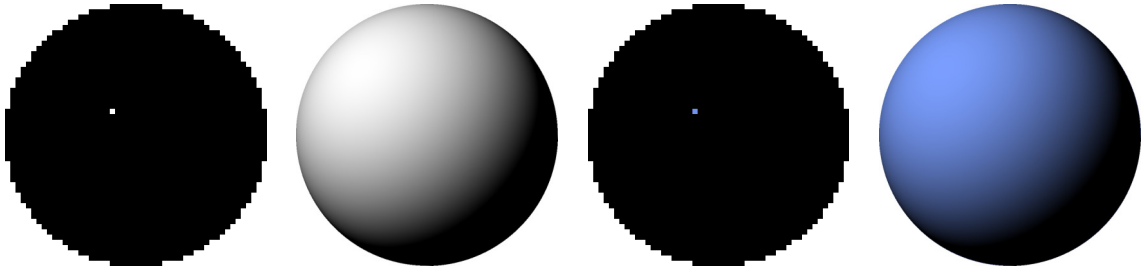
Figure 2.7: Illustration depicting radiance and shading. The images in (a) and (c) each showcase a radiance map, where (a) is the grayscale version of (c). The images in (b) and (d) display the shading, *i.e.*, the evaluation of Equation (2.38). In the shading images, the normals correspond to a hemisphere, and the incoming radiance aligns with the radiance maps displayed in (a) and (c).

include a lack of inter-reflections and the absence of cast shadows. It is important to note that this assumption is not universally applicable, although there are real-world situations in which the spatial independence outlined in Equation (2.37) holds true. For instance, a class of objects which do not experience cast shadows are convex objects, making it theoretically feasible for a convex scene to conform to this assumption.
Assuming this, the shading integral that is the focus of this section simplifies from (2.27) to the form

$$S(\mathbf{x}) = \int_{\mathbb{S}^2} L_\mathrm{i}(\omega_\mathrm{i}) \max(0, \langle \mathbf{n}(\mathbf{x}), \omega_\mathrm{i} \rangle) \, \mathrm{d}\omega_\mathrm{i}, \tag{2.38}$$

see also Figure 2.7 for a visualization of exemplary radiances $L_\mathrm{i}$, as well as the corresponding shadings of a white hemisphere. Several algorithms that leverage the distant light assumption by solving the shading equation (2.38) are categorized as environment map algorithms, as discussed in [191]. The simplification of the lighting conditions provides the benefit of reducing the complexity of the shading term $S$, which may avoid the need for the evaluation of the intricate integral set forth in Equation (2.38). Nevertheless, simplifications can sometimes detract from the realism of the model, as certain facets of the real-world are not accurately represented, such as the absence of cast shadows as previously described. In addition to their limitations, we will now delve into two commonly employed assumptions pertaining to the lighting model and how they can greatly facilitate the simplification of the shading integral.

[14]Environment map taken from `http://www.hdrlabs.com/sibl/archive.html`, accessed on 29th of March, 2019 at 3.39PM.



(a) Grayscale radiance $L_\mathrm{i}$ showing directional lighting. Note the white point within the sphere, indicating the presence of non-zero incoming radiance.

(b) Shading $S$ of a white hemisphere using the radiance map shown in (a).

(c) RGB radiance $L_\mathrm{i}$ showing directional lighting. Note the blueish point within the sphere, indicating the presence of non-zero incoming radiance.

(d) Shading $S$ of a white hemisphere using the radiance map shown in (c).

Figure 2.8: Illustration depicting directional light and shading. The images in (a) and (c) each showcase a radiance map of directional light. The images in (b) and (d) display the shading under directional light, i.e., the evaluation of Equation (2.42). In the shading images, the normals correspond to a hemisphere, and the incoming radiance aligns with the radiance maps displayed in (a) and (c).

#### 2.3.3.1 Directional Lighting

In the case of *directional lighting*, it is assumed that only a single lighting direction, $\omega_\mathrm{i}^\mathrm{dir}$, is present, with incoming radiance that does not diminish to zero,

$$L_\mathrm{i}^\mathrm{dir}(\omega_\mathrm{i}) = \begin{cases} L_\mathrm{i}(\omega_\mathrm{i}), & \text{if } \omega_\mathrm{i} = \omega_\mathrm{i}^\mathrm{dir}, \\ 0, & \text{if } \omega_\mathrm{i} \neq \omega_\mathrm{i}^\mathrm{dir}. \end{cases} \tag{2.39}$$

Directional lighting is a frequently employed assumption when an object is situated in a large, dark environment, where the distance between the object and the light source is significantly greater than the diameter exhibited by the object. By substituting (2.39) into (2.38), the shading integral simplifies to a single evaluation,

$$\begin{aligned} S(\mathbf{x}) &= \int_{\mathbb{S}^2} L_\mathrm{i}^\mathrm{dir}(\omega_\mathrm{i}) \max(0, \langle \mathbf{n}(\mathbf{x}), \omega_\mathrm{i} \rangle) \, \mathrm{d}\omega_\mathrm{i} && (2.40) \\ &= L_\mathrm{i}\big(\omega_\mathrm{i}^\mathrm{dir}\big) \max\big(0, \big\langle \mathbf{n}(\mathbf{x}), \omega_\mathrm{i}^\mathrm{dir} \big\rangle\big) && (2.41) \\ &= \max\big(0, \big\langle \mathbf{n}(\mathbf{x}), \mathbf{l}^\mathrm{dir} \big\rangle\big), && (2.42) \end{aligned}$$

where in the final step, the light direction, $\omega_\mathrm{i}^\mathrm{dir}$, is scaled by the light intensity, $L_\mathrm{i}\big(\omega_\mathrm{i}^\mathrm{dir}\big)$, resulting in a single directional light vector, $\mathbf{l}^\mathrm{dir} = L_\mathrm{i}\big(\omega_\mathrm{i}^\mathrm{dir}\big)\omega_\mathrm{i}^\mathrm{dir} \in \mathbb{R}^3$. Exemplary directional-lighting-based radiances $L_\mathrm{i}^\mathrm{dir}$ and shadings are shown in Figure 2.8.

The rendering equation governing the radiance at a non-emissive point $\mathbf{x}$ with Lambertian reflectance, and under directional lighting can be expressed as

$$L_{\mathrm{o}}^{\mathrm{dir}}(\mathbf{x}) = f_{\mathrm{BRDF}}^{\mathrm{d}}(\mathbf{x}) \max\big(0, \langle \mathbf{n}(\mathbf{x}), \mathbf{l}^{\mathrm{dir}} \rangle\big) \tag{2.43}$$

$$= \rho(\mathbf{x}) \max\big(0, \langle \mathbf{n}(\mathbf{x}), \mathbf{l}^{\mathrm{dir}} \rangle\big). \tag{2.44}$$

By employing these simplifications, an image can be rendered in a straightforward manner. The rendering process can be succinctly described through the use of an albedo map $\rho : \Omega \to [0, 1]$, a lighting vector $\mathbf{l}^{\mathrm{dir}} \in \mathbb{R}^3$, and a normal map $\mathbf{n} : \Omega \to \mathbb{S}^2$. Once these components are provided, the image can be rendered as

$$I(\mathbf{p}) = \rho(\mathbf{p}) \max\big(0, \langle \mathbf{n}(\mathbf{p}), \mathbf{l}^{\mathrm{dir}} \rangle\big). \tag{2.45}$$

While the simplicity and ease of computation of this model is advantageous, it is limited in that it represents a fairly controlled laboratory setup, assuming only a single distant light direction in a dark environment. In the following section, we introduce a more sophisticated model based on the spherical harmonics (SH) framework, which enables the simulation of natural lighting scenarios.

#### 2.3.3.2 Natural Lighting

Compared to directional lighting, the illumination scenario that we discuss here is more realistic. We consider a non-emissive and Lambertian surface that is illuminated with distant lighting. Our aim is to directly simplify Equation (2.38). This scenario is referred to as *natural illumination* or *general illumination*, and examples of it include a non-dark room with multiple light sources [72] or an outdoor environment with light from the sky on a cloudy day[15] [105]. In this section, we largely follow the works of [21, 191] for the relationship between shading and SH, and [28, 49] to derive the SH that map to the real numbers $\mathbb{R}$ from the SH that map to the complex numbers $\mathbb{C}$.
In order to facilitate the shading process outlined in (2.38), it is possible to interpret it as a convolution of the incoming radiance function $L_{\mathrm{i}} : \mathbb{S}^2 \to \mathbb{R}_0^+$ with a kernel function $K : \mathbb{S}^2 \to \mathbb{R}_0^+$,

$$K(\omega_{\mathrm{i}}) = \max(0, \langle \mathbf{n}(\mathbf{x}), \omega_{\mathrm{i}} \rangle). \tag{2.46}$$

Both of these functions are non-negative over the surface of the sphere $\mathbb{S}^2$, and as such can be represented using an orthonormal basis on this surface. One widely-used basis are the spherical harmonics, which are defined using the *associated Legendre polynomials*

[15] Cast shadows resulting from a sunny day are not modeled with (2.38) due to the spatial independence of the incoming radiance.

$P_n^m : [-1, 1] \to \mathbb{R}$,

$$P_n^m(x) = \frac{(1-x^2)^{\frac{m}{2}}}{2^n n!} \frac{\mathrm{d}^{n+m}}{\mathrm{d}x^{n+m}} \left[ \left(x^2 - 1\right)^n \right] \tag{2.47}$$

and a normalization factor $N_n^m$,

$$N_n^m = \sqrt{\frac{2n+1}{4\pi} \frac{(n-m)!}{(n+m)!}}. \tag{2.48}$$

From this, we can define the *spherical harmonics (SH)* function of degree $n$ and order $m$, denoted as $Y_n^m : \mathbb{S}^2 \to \mathbb{C}$,

$$Y_n^m(\theta, \phi) = N_n^m P_n^m(\cos\theta) \exp(im\phi), \tag{2.49}$$

where we "abusively" use the spherical coordinates $(\theta, \phi)$ to describe elements on $\mathbb{S}^2$. Specifically, $\theta \in [0, \pi]$ denotes the *polar angle*, also known as the *colatitude*, while $\phi \in [0, 2\pi)$ denotes the *azimuth*, or *longitude*.

Since the SH are complex-valued functions that map to the complex plane, they are not directly suitable for computing the real-valued shading term in Equation (2.38). To address this, we can derive real-valued functions from the SH functions by combining complex conjugate functions of $Y_n^m$. The resulting real SH functions are given by $Y_{nm} : \mathbb{S}^2 \to \mathbb{R}$,

$$Y_{nm}(\theta, \phi) = \begin{cases} N_n^m P_n^m(\cos\theta)\sqrt{2}\cos m\phi, & \text{if } m > 0, \\ N_n^0 P_n^0(\cos\theta), & \text{if } m = 0, \\ N_n^{|m|} P_n^{|m|}(\cos\theta)\sqrt{2}\sin|m|\phi, & \text{if } m < 0. \end{cases} \tag{2.50}$$

For the sake of conciseness, we do not present the algebraic derivation here, but instead refer interested readers to the work of Blanco *et al.* [28].
The expression in (2.50) represents an orthonormal basis on the sphere [28], which provides a convenient means to represent the incoming radiance $L_\mathrm{i} : \mathbb{S}^2 \to \mathbb{R}_0^+$ and the convolution kernel $K : \mathbb{S}^2 \to \mathbb{R}_0^+$ as an infinite linear combination of the SH basis functions,

$$L_\mathrm{i}(\omega_\mathrm{i}) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} l_{nm} Y_{nm}(\omega_\mathrm{i}) \tag{2.51}$$

$$K(\omega_\mathrm{i}) = \sum_{n=0}^{\infty} k_n Y_{n0}(\omega_\mathrm{i}), \tag{2.52}$$

where the coefficients of the radiance, $l_{nm}$ can be interpreted as the "lighting" vector.

The coefficients of the kernel depend solely on the SH functions with $m = 0$, as the kernel exhibits symmetry around $\mathbf{n}$ (the north pole). Notably, these coefficients $k_n$ can be efficiently computed thanks to the kernel's particular structure,

$$k_n = \sqrt{\frac{4\pi}{2n+1}} \cdot \begin{cases} \frac{\sqrt{\pi}}{2}, & \text{if } n = 0, \\ \sqrt{\frac{\pi}{3}}, & \text{if } n = 1 \\ (-1)^{\frac{n}{2}+1} \frac{\sqrt{(2n+1)\pi}}{2^n (n-1)(n+2)} \binom{n}{\frac{n}{2}}, & \text{if } n \geq 2, \text{ even}, \\ 0, & \text{if } n \geq 2, \text{ odd}. \end{cases} \tag{2.53}$$

We can apply the Funk-Hecke theorem [80] to demonstrate that the substitution of (2.51) and (2.52) into (2.38) is tantamount to multiplication of the coefficients,

$$S(\mathbf{x}) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} l_{nm} k_n Y_{nm}(\mathbf{n}(\mathbf{x})). \tag{2.54}$$

The evaluation of the SH functions is now performed w.r.t. the normal $\mathbf{n}$, since the integral in (2.38) that we express in terms of SH functions can be considered as a function of the surface normal. It can not be interpreted as a function of the incoming direction $\omega_\mathrm{i}$, as this is the integration variable. This simplification is advantageous, as it enables us to compute the integral in (2.38) with a summation. To compute the (real) SH functions up to a high degree, it is recommended to use a numerically stable method based on the recursive computation of the associated Legendre polynomials [28, 181], combined with widely used trigonometric identities such as the recursive *multiple-angle formulae* for $\cos m\phi$ and $\sin|m|\phi$.

Regrettably, the summation in (2.54) extends infinitely, but the first two harmonic degrees are sufficient to capture most of the energy. In fact, a first- and second-degree approximation capture $75\%$ and $97.96\%$, respectively, of the non-negative light [21]. Therefore, it is justified to truncate the summation at $n = 2$,

$$\sum_{n=0}^{\infty} \sum_{m=-n}^{n} l_{nm} k_n Y_{nm}(\mathbf{n}(\mathbf{x})) \approx \sum_{n=0}^{2} \sum_{m=-n}^{n} l_{nm} k_n Y_{nm}(\mathbf{n}(\mathbf{x})). \tag{2.55}$$

One may observe that the real SH functions presented in (2.50) are defined in terms of spherical coordinates, while our incoming radiance (2.37) and kernel (2.46) are defined on unit vectors. To address this discrepancy, it comes convenient to express the real SH in terms of spatial coordinates $\mathbf{x} = (x, y, z)^\top \in \mathbb{S}^2$. This can be achieved through the use of the aforementioned recursive approach and the spherical to Cartesian coordinate transformation, $(\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta)^\top = (x, y, z)^\top \in \mathbb{S}^2$. For our purposes, it

suffices to specify the first nine real SH up to second degree in Cartesian coordinates for a given point $\mathbf{x} \in \mathbb{S}^2$:

$$
\begin{array}{lllll}
 & n=0 & n=1 & n=2 & \\
m=-2 & & & Y_{2-2}(\mathbf{x}) = \sqrt{\frac{15}{4\pi}}xy & (2.56)\\
m=-1 & & Y_{1-1}(\mathbf{x}) = \sqrt{\frac{3}{4\pi}}y & Y_{2-1}(\mathbf{x}) = \sqrt{\frac{15}{4\pi}}yz & (2.57)\\
m=0 & Y_{00}(\mathbf{x}) = \frac{1}{\sqrt{4\pi}} & Y_{10}(\mathbf{x}) = \sqrt{\frac{3}{4\pi}}z & Y_{20}(\mathbf{x}) = \sqrt{\frac{5}{16\pi}}\left(3z^2-1\right) & (2.58)\\
m=1 & & Y_{11}(\mathbf{x}) = \sqrt{\frac{3}{4\pi}}x & Y_{21}(\mathbf{x}) = \sqrt{\frac{15}{4\pi}}xz & (2.59)\\
m=2 & & & Y_{22}(\mathbf{x}) = \sqrt{\frac{15}{16\pi}}\left(x^2-y^2\right) & (2.60)
\end{array}
$$

The weights of the corresponding kernel $k_n$ up to the second degree can be expressed as

$$
\begin{array}{lll}
n=0 & n=1 & n=2 \\
k_0 = \pi & k_1 = \frac{2\pi}{3} & k_2 = \frac{\pi}{4}.
\end{array} \tag{2.61}
$$

For more details, we refer the interested reader to [21].
SH of both first and second degree are frequently employed in real-world scenarios due to their ability to capture complex phenomena while retaining a manageable level of complexity. Numerous works in the field have utilized these functions for a variety of applications [3, 5, 6, 7, 8, 9, 140, 149, 150, 151, 164, 188, 263]. The first-degree approximation ($n = 1$) has the desirable property of being linear w.r.t. the surface normal. However, this property does not hold for the second-degree approximation ($n = 2$), as can be seen in equations (2.56) to (2.60), due to the presence of non-linear terms.
Given the Lambertian reflectance assumption, we may insert the relevant terms into the rendering equation. Specifically, we substitute (2.55) into (2.54) and the distant lighting shading (2.38), thereby arriving at,

$$
L_{\mathrm{o}}^{\mathrm{SH}}(\mathbf{x}) = \rho(\mathbf{x})\left\langle \mathbf{l}^{\mathrm{SH}}, \mathbf{Y}_n(\mathbf{n}(\mathbf{x}))\right\rangle, \tag{2.62}
$$

where $\mathbf{l}^{\mathrm{SH}} \in \mathbb{R}^{(n+1)^2}$ are the stacked "light" coefficients, $\{l_{nm}\}_{nm}$, and $\mathbf{Y}_n(\mathbf{n}(\mathbf{x})) \in \mathbb{R}^{(n+1)^2}$ are the stacked kernel weighted SH basis functions evaluated at the surface normal, $\{k_n Y_{nm}(\mathbf{n}(\mathbf{x}))\}_{nm}$. It is a common practice to omit the weights that appear in front

(a) Image of a white hemisphere rendered under first-degree spherical harmonics lighting, $\mathbf{Y}_1$.

(b) Image of a white hemisphere rendered under second-degree spherical harmonics lighting, $\mathbf{Y}_2$.

Figure 2.9: Illustration of spherical harmonics lighting. These rendered images are generated by applying Equation (2.63) with a white albedo and a hemisphere geometry. The light coefficients $\mathbf{l}^{\text{SH}}$ approximate the incoming radiance depicted in Figure 2.7(c), hence the images visualized here approximate the image shown in Figure 2.7(d).

of the SH functions (see (2.56) – (2.60)) as well as the $k_n$'s since they can be incorporated into the lighting vector, $\mathbf{l}^{\text{SH}}$, as simple constants. When calibrating or optimizing $\mathbf{l}^{\text{SH}}$, multiplication with a constant is immaterial [6].
Similar to the case of directional lighting, image rendering can be performed in a straightforward manner. The rendering process can be concisely described by utilizing an albedo map $\rho : \Omega \to [0, 1]$, a lighting vector $\mathbf{l}^{\text{SH}} \in \mathbb{R}^{(n+1)^2}$, and a normal map $\mathbf{n} : \Omega \to \mathbb{S}^2$, as

$$I(\mathbf{p}) = \rho(\mathbf{p})\left\langle \mathbf{l}^{\text{SH}}, \mathbf{Y}_n(\mathbf{n}(\mathbf{p}))\right\rangle. \tag{2.63}$$

To provide visual illustration in Figure 2.9, two example images are rendered utilizing Equation (2.63).

In this section, we have derived two distinct lighting scenarios: the directional scenario as described in (2.45) and the natural scenario as outlined in (2.63). Although these formulations may appear similar at first glance, there are subtle yet significant differences that are worthy of mention.
The SH functions are utilized for the approximation of the radiance (2.37), the convolution kernel (2.46), and eventually the shading (2.38). It should be noted that the direc-

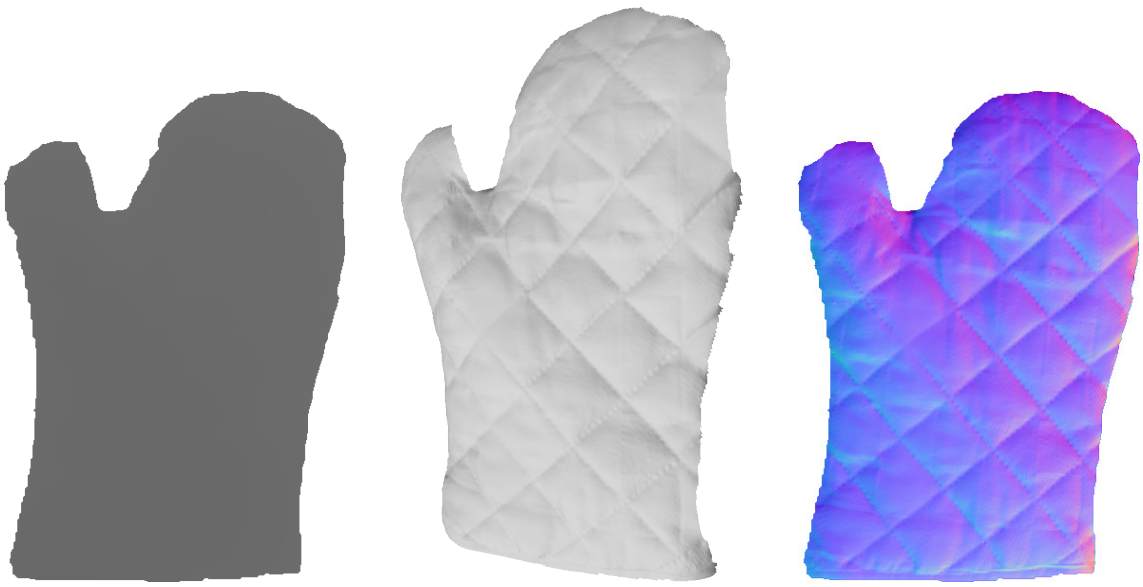
tional lighting scenario, imposes a strong assumption on the incoming radiance, as it is essentially a dirac function, as shown in (2.39). Its resulting effect can be demonstrated via using two incoming radiance maps presented in the previous figures. Figure 2.7(c) showcases a smooth outdoor scenario[16], specifically a road in Monument Valley, USA, which can be accurately approximated by employing SH functions, as depicted in Figure 2.9. Furthermore, Figure 2.8(c) illustrates a directional light scenario using a Dirac function. Although the SH functions could potentially approximate the directional lighting, they may be slow to converge, necessitating the use of numerous basis functions. *E.g.*, when employing a first-degree approximation, the natural and directional lighting scenarios vary solely in terms of the $\max$ operator and the $0$-th degree constant SH function, also referred to as the *ambient light*. Despite their apparent similarity in this instance, the accuracy of first-degree SH functions in approximating a dirac function may be significantly compromised [72].
Furthermore, the $\max$ operator appearing in (2.38) plays a crucial role in preventing radiance values from becoming negative. While this operator is still present in the directional light simplification in (2.45), it is not strictly invoked in the natural light approximation based on SH functions in (2.63). As a consequence, low-degree SH approximations may lead to negative radiance values, a fact that should be kept in mind when generating data according to (2.63) and storing it as images[17], as these negative values will be clamped to $0$ or wrapped around $255$.

Both expressions in (2.45) and (2.63) share a common dependency on the surface normal $\mathbf{n}$ at the point $\mathbf{x}$ or the corresponding pixel $\mathbf{p}$. However, as discussed in Section 2.2, when using RGB-D cameras, depth maps are available instead of surface normals. While a simple image can be rendered using a normal and albedo map with some lighting vector, as shown in Equations (2.45) and (2.63), it may not be immediately clear where and why depth maps should be incorporated. In the subsequent section, we showcase the advantageous impact of the inclusion of depth maps in comparison to normal maps.

### 2.3.4 Geometry

In this section, we delve into the geometric considerations of the rendering equation. Typically, when rendering scenes from multiple viewpoints, a mesh or a signed distance function (SDF) is utilized. This is the case in Chapter 8, where a mesh was employed as the underlying geometric structure. However, in the upcoming Chapters 5 – 7, only a single object is rendered from a sole viewpoint. Therefore, it suffices to represent the object's geometry using a single image, such as a depth or a normal map. To this end,

[16] In this situation, it is essentially an environment map
[17] We refer to standard image formats like png, jpg, et cetera.



(a) Depth map (b) 3D shape (c) Surface normals

Figure 2.10: Illustration depicting the relationship between depth, shape, and surface normals using the oven mitt object from Figure 2.3. Although the depth map alone may lack intricate details, projecting it into 3D and visualizing its shape as a mesh effectively reveals the fine structure of the oven mitt. This fine structure is also evident in the normal map, which is derived directly from the depth map in (a).

further elaboration is required to discuss the relationship between the two parameterizations and to examine their respective advantages and disadvantages. In detail, we will describe how to calculate normals from a depth map depending on its underlying projection, and analyze the trade-offs between using a depth versus a normal parameterization.

Surface normals and depth maps are intimately related, as one can be derived from the other by computing its gradient, *i.e.*, we can directly parameterize the surface normals through depth. Therefore, when evaluating equations (2.45) or (2.63), we still compute them based on surface normals, but we deduce these normals from their corresponding depth maps. Figure 2.10 showcases a depth map, its 3D shape, and the corresponding normal map, providing a visual representation of their interconnected relationship, which will be further explored and deduced mathematically in the following. As discussed in Sections 2.1 and 2.2, a depth map, denoted by $z : \Omega \to \mathbb{R}$, is a 2D grayscale image, where each pixel position, $\mathbf{p} \in \Omega$, describes the distance, $z(\mathbf{p})$, to the corresponding 3D point, $\mathbf{x} = \Pi^{-1}(\mathbf{p}, z(\mathbf{p}))$. Note that the camera model described by $\Pi$ must be consistent with the parameterization of the depth map $z$, *i.e.*, an orthographic or perspective camera must be used for an orthographic or perspective depth map, re-

spectively. Thus, we can reformulate a normal vector, $\mathbf{n}$, in terms of its underlying depth map, $z$, at pixel $\mathbf{p} = (u, v)^\top$ as

$$\mathbf{n}(\mathbf{x}) = \mathbf{n}\big(\Pi^{-1}(\mathbf{p}, z(\mathbf{p}))\big). \tag{2.64}$$

Equation (2.64) demonstrates that the camera model, $\Pi$, plays a critical role in determining the surface normal. Specifically, the unit normal, now w.r.t. $\mathbf{p}$ and $z$ in direction of the principal axis, can be expressed as the normalized cross product of the partial derivatives,

$$\mathbf{n}[z](\mathbf{p}) = \mathrm{normalize}\left(\frac{\partial \Pi^{-1}(\mathbf{p}, z(\mathbf{p}))}{\partial u} \times \frac{\partial \Pi^{-1}(\mathbf{p}, z(\mathbf{p}))}{\partial v}\right), \tag{2.65}$$

where we obtain the normalized vector $\mathrm{normalize}(\mathbf{x})$ from the original vector $\mathbf{x}$ by dividing it by its $L_2$-norm, $\mathrm{normalize}(\mathbf{x}) = \frac{\mathbf{x}}{\|\mathbf{x}\|_2}$. One can deduce the partial derivatives of the inverse projection w.r.t. its pixel positions using the chain rule,

$$\frac{\partial \Pi^{-1}(\mathbf{p}, z(\mathbf{p}))}{\partial u} = \frac{\partial \Pi^{-1}}{\partial u} + \frac{\partial \Pi^{-1}}{\partial z}\frac{\partial z}{\partial u} \tag{2.66}$$

$$\frac{\partial \Pi^{-1}(\mathbf{p}, z(\mathbf{p}))}{\partial v} = \frac{\partial \Pi^{-1}}{\partial v} + \frac{\partial \Pi^{-1}}{\partial z}\frac{\partial z}{\partial v}. \tag{2.67}$$

We introduce the notation for partial derivatives of the depth map w.r.t. its pixel positions as $z_u = \frac{\partial z}{\partial u}$ and $z_v = \frac{\partial z}{\partial v}$, which enable us to express the gradient of $z$ as $\nabla z = (z_u, z_v)^\top$. The computation of these partial derivatives typically involves a discrete stencil such as forward differences, or backpropagation in the case of depth representation using a neural network. In this thesis, unless otherwise specified, we employ forward differences with Neumann boundary conditions for computing the gradient in the image plane. Furthermore, the partial derivatives $\frac{\partial \Pi^{-1}}{\partial u}$, $\frac{\partial \Pi^{-1}}{\partial v}$, and $\frac{\partial \Pi^{-1}}{\partial z}$ depend on the camera model, as discussed in Section 2.1 for orthographic ($\Pi_\mathrm{o}$) and perspective projection ($\Pi_\mathrm{p}$). We explicitly state these partial derivatives of both projection types along with the final result of the computed normal (2.65) in the following.

**Orthographic Projection.** For the case of orthographic projection, the partial derivatives can be expressed in a straightforward manner,

$$\frac{\partial \Pi_\mathrm{o}^{-1}}{\partial u} = \begin{pmatrix}1\\0\\0\end{pmatrix}, \qquad \frac{\partial \Pi_\mathrm{o}^{-1}}{\partial v} = \begin{pmatrix}0\\1\\0\end{pmatrix}, \qquad \frac{\partial \Pi_\mathrm{o}^{-1}}{\partial z} = \begin{pmatrix}0\\0\\1\end{pmatrix}. \tag{2.68}$$

The normal in the orthographic case can be computed by combining these partial derivatives of the inverse projection and the gradient of the depth map. Therefore, the unit normal vector pointing upward can be expressed as,

$$\mathbf{n}[z](\mathbf{p}) = \text{normalize} \begin{pmatrix} -z_u \\ -z_v \\ 1 \end{pmatrix} \tag{2.69}$$

$$= \text{normalize} \begin{pmatrix} -\nabla z(\mathbf{p}) \\ 1 \end{pmatrix}. \tag{2.70}$$

**Perspective Projection.** The perspective case entails a slightly more intricate computation of the partial derivatives,

$$\frac{\partial \Pi_{\text{p}}^{\text{-1}}}{\partial u} = \begin{pmatrix} f_x^{-1} z(\mathbf{p}) \\ 0 \\ 0 \end{pmatrix}, \quad \frac{\partial \Pi_{\text{p}}^{\text{-1}}}{\partial v} = \begin{pmatrix} 0 \\ f_y^{-1} z(\mathbf{p}) \\ 0 \end{pmatrix}, \quad \frac{\partial \Pi_{\text{p}}^{\text{-1}}}{\partial z} = \begin{pmatrix} f_x^{-1}(u - c_x) \\ f_y^{-1}(v - c_y) \\ 1 \end{pmatrix}. \tag{2.71}$$

Upon substituting (2.71) associated with the perspective camera model into Equation (2.65) and simplifying the expression, we obtain

$$\mathbf{n}[z](\mathbf{p}) = \text{normalize} \left( \frac{z(\mathbf{p})}{f_x f_y} \begin{pmatrix} -f_x z_u(\mathbf{p}) \\ -f_y z_v(\mathbf{p}) \\ z + \left\langle \begin{pmatrix} v - c_x \\ u - c_y \end{pmatrix}, \begin{pmatrix} z_u(\mathbf{p}) \\ z_v(\mathbf{p}) \end{pmatrix} \right\rangle \end{pmatrix} \right) \tag{2.72}$$

$$= \text{normalize} \begin{pmatrix} -\nabla_f z(\mathbf{p}) \\ z + \langle \mathbf{p} - c, \nabla z(\mathbf{p}) \rangle \end{pmatrix}. \tag{2.73}$$

In the final step, we have disregarded the constant factor $\frac{z(\mathbf{p})}{f_x f_y}$, as it does not impact the result. Additionally, we have utilized the scaled gradient operator $\nabla_f$ and the principal point $c$ as defined in (2.14) [79].

In both the orthographic and perspective cases, the normal vector is oriented upward in the direction of the principal axis. However, in most applications, it is preferable for the normal to point towards the camera. To achieve this, it is common practice to invert the normal vector by multiplying it with $-1$.

We have established the means by which a normal can be calculated based on a given depth map. However, the question of its significance still looms large. After all, a normal map suffices to render images, as shown in (2.45) and (2.63). The rationale behind a depth-based representation is two-fold. First, the normal map is inherently integrable

if represented using depth. In other words, every depth map infers a normal map, but not every normal map has an underlying depth map. A normal map is said to be *integrable* if a surface can be described with it. The forward process of computing a normal map as shown in (2.70) and (2.73) is straightforward when a depth map is provided. The backward process, known as *normal integration*, is not as trivial and is beyond the scope of this thesis. Interested readers may refer to [184, 185] for more information. In many 3D reconstruction tasks, the desired outcome is a surface. However, if the estimated normals are not integrable, then fitting a depth map to the normal field can result in undesired artifacts and surfaces that are far from the genuine geometry. Directly estimating over depth as a trade-off for avoiding such issues results in normalization issues that introduce non-convexity when optimizing for depth. However, such issues can be addressed via more sophisticated optimization schemes [1, 3, 4, 5, 6, 7, 8, 149, 182, 183, 188, 189, 190] or algebraic reformulations [78, 133, 143, 144, 145, 146, 147, 148, 187, 211]. The second reason for representing a surface with a depth map is rooted in the assumption that an initial geometry is available, for instance, from an RGB-D camera's depth sensor [3, 5, 7, 8, 86, 164, 188, 239, 250] or a generic minimal surface of specific volume [6, 167]. To obtain a refined and superior version of the initial depth map, it can be beneficial to directly optimize over the depth map instead of employing a two-step approach consisting of first optimizing the normal field and then integrating the potentially non-integrable normal field to infer depth. In this two-step process, the final depth map is determined only by the estimated normal map and is not influenced by the initial step.

With the image formation models presented in the previous and current sections, it is now possible to render images of diffuse objects in a straightforward manner without explicitly evaluating the shading integral (2.38). Specifically, we can combine the concepts discussed in Section 2.3.3 and Section 2.3.4 as desired, resulting in four possible *(projection, lighting)* pairs when representing surfaces in terms of depth: *(orthographic, directional)*, *(orthographic, SH)*, *(perspective, directional)*, and *(perspective, SH)*. Two of these combinations have been employed in the contributed papers of this thesis. A perspective projection with SH lighting was utilized in an SfS problem in Chapter 5 and in a PS problem in Chapter 6, whereas an orthographic projection with directional lighting was utilized in a PS problem in Chapter 7.
Although the forward process can be readily performed by evaluating the rendering equation or its approximations based on the scene's assets, such as geometry, material, and illumination, the inverse process of recovering one or possibly multiple assets from one or multiple images is a critical branch of 3D reconstruction. In this context, we will now take a close look at inverting the rendering equation, as well as the topics of SfS and PS.

## 2.4 Inverting the Rendering Equation

The focus of our discussion now shifts towards the inversion of the rendering equation. In the fields of computer graphics and computer vision, it is a crucial question to recover geometry, material, and/or illumination information from a given set of images. In this thesis, we will investigate a universal formulation of the inversion process as an inverse problem. Subsequently, we provide an in-depth analysis of two prominent problems, namely Shape-from-Shading (SfS) and photometric stereo (PS).

The process of inverting the rendering equation is commonly referred to as *inverse rendering*. Depending on the specific setup, it can be well-posed [1, 170] or ill-posed [14, 20, 88, 171, 252]. Furthermore, non-convexities [6, 256] and non-differentiabilities [126, 138, 255] can often arise, making inverse rendering a challenging problem in general.
Let us consider the problem of recovering a set of parameters $\mathcal{X}$ that represent various aspects of the scene, including geometric, reflectance, and lighting properties. It should be noted that in some cases, $\mathcal{X}$ may also include other attributes, such as camera parameters, as observed in recent studies [31, 140, 232]. However, our research focuses exclusively on the recovery of geometry, reflectance, and lighting.
Given a set of $N \in \mathbb{N}$ images $\mathcal{I} = \{I_i\}_{i=1,\dots,N}$, where $I_i$ represents the $i$-th image, the inverse rendering problem can be formulated mathematically as the optimization of an objective function $f$,

$$\min_{\mathcal{X}} f(\mathcal{I}; \mathcal{X}) = E_{\text{data}}(\mathcal{I}; \mathcal{X}) + E_{\text{reg}}(\mathcal{X}), \tag{2.74}$$

consisting of a data term $E_{\text{data}}$ and a regularization term $E_{\text{reg}}$. The data term in the formulation of the inverse rendering problem serves to quantify the difference between the input and rendered images through the use of a residual, such as the photometric difference. On the other hand, the regularization term typically operates solely on the optimization parameters $\mathcal{X}$. This is done to address the presence of noise and ensure the validity of the solution, as well as to limit the search space and prevent ambiguities.
To emphasize the dependence on $\mathcal{X}$ in the subsequent sections, we shall express the outgoing radiance $L_{\text{o}}$ as a function of its optimization parameters, rather than the pixel or point positions.
In the upcoming sections, we shall analyze the problems of SfS and PS that are aimed at resolving particular types of (2.74). We will investigate their individual assumptions, challenges, and various techniques to overcome some of their respective obstacles.

(a) Input image. (b) Output normal map.

Figure 2.11: Illustration of the SfS problem. From a single input image, as depicted in (a) infer the corresponding geometry, represented here by a normal map in (b)[18].

### 2.4.1 Shape-from-Shading

Solving the problem of inferring the shape of an object from shading clues using a single image dates back to the 1970 Ph.D. thesis of Berthold Horn [92] and is commonly known as the method of *Shape-from-Shading (SfS)*. An illustration of the SfS problem is shown in Figure 2.11. This method involves formulating an optimization problem w.r.t. $\mathcal{X}$ based on the inverse rendering formulation shown in (2.74),

$$\min_{\mathcal{X}} \|I - L_{\mathrm{o}}(\mathcal{X})\|_2^2 + E_{\mathrm{reg}}(\mathcal{X}), \tag{2.75}$$

where, depending on the task at hand, the parameter $\mathcal{X}$ can take the form of either a depth map $z : \Omega \to \mathbb{R}$ or a normal map $\mathbf{n} : \Omega \to \mathbb{S}^2$.
Regrettably, the SfS problem is severely ill-posed, as illustrated by Adelson and Pentland's workshop metaphor [14]. For instance, a painter may describe an image as a flat shape illuminated uniformly but painted in a complex manner, while a sculptor may describe an image as a white and frontally-lit surface with a complex geometry. Similarly, a gaffer may explain an image as a white planar surface illuminated in a complex manner.

[18]Geometry taken from `https://www.thingiverse.com/thing:897412/remixes`, accessed on 7th of July, 2023 at 4.01PM.

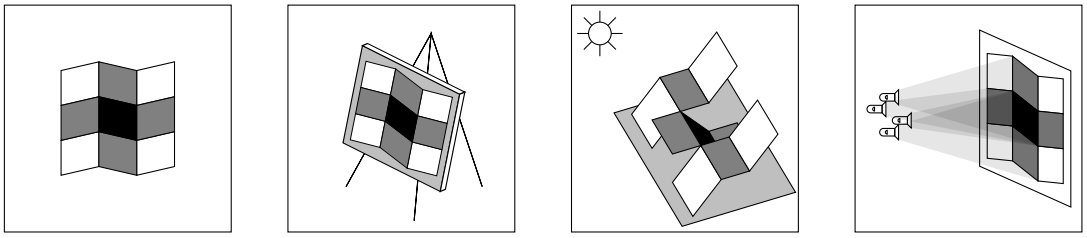



Figure 2.12: Illustration of the workshop metaphor by Adelson and Pentland [14]. Each artist provides a distinct interpretation of the same image. The painter emphasizes complex color, flat geometry, and uniform illumination in their description. The sculptor focuses on complex geometry, white color, and uniform illumination. The gaffer highlights complex illumination, flat geometry, and white color. This metaphor vividly demonstrates the inherent challenges and ill-posed nature of recovering scene properties from a single image.

An illustration of the workshop metaphor [14] is provided in Figure 2.12, which depicts the diverse ways in which a single image can be explained by different artists. Let us closely examine the sculptor's explanation. The assumption is that the scene depicted in the image is white and frontally-lit. The former can be realized by setting the albedo to one, denoted as $\rho = 1$, while the latter can be achieved by setting the directional lighting vector $\mathbf{l}^{\text{dir}} = (0, 0, -1)^\top$. By incorporating these assumptions into the image formation model, specifically the directional lighting model shown in Equation (2.45), and if we further assume orthographic projection (2.70), the image intensity can be described as

$$I(\mathbf{p}) = \frac{1}{\sqrt{\|\nabla z(\mathbf{p})\|^2 + 1}}. \tag{2.76}$$

Note that we have adjusted the directions of both the normal and the light vector to point away from the surface, towards the camera. The simplified case in (2.76) reveals that the image intensity solely relies on the gradient of the depth, which corresponds to the local change of depth. Therefore, given the image intensities and the desire to obtain the depth map, (2.76) can be employed to solve for $z$ [35]. This leads to the well-known *Eikonal equation*[19] as follows

$$\|\nabla z(\mathbf{p})\| = \sqrt{\frac{1}{I(\mathbf{p})^2} - 1}. \tag{2.77}$$

[19]The Eikonal equation is a non-linear first-order partial differential equation (PDE) that endeavors to solve for $u$ in the expression $\|\nabla u\| = f$, where $f$ is a given positive function. By setting $u$ as $z$ and $f$ as $\sqrt{\frac{1}{I^2} - 1}$, we obtain (2.77).

Hence, even in the simplified case of the sculptor where both the albedo and illumination are known, the depth cannot be deduced with certainty. We only have information about the magnitude of the gradient, without knowledge of its direction or sign. Solving the Eikonal equation, as depicted in Equation (2.77), is beyond the scope of this thesis. However, this highlights the significant ill-posedness of the classical SfS problem, as even under strong assumptions, the problem remains ambiguous. Interested readers may refer to works that discuss mathematical solutions of the PDE shown in Equation (2.77) in terms of viscosity, such as [34, 52, 66, 131, 180, 195], as well as two surveys on SfS [60, 257].
We are primarily concerned with solving the variational problem stated in (2.75) [71, 94, 228]. However, we aim to enhance the realism of our SfS approach by considering perspective projection [34, 104, 180, 219] and natural illumination [95, 102, 169, 193]. In order to increase the robustness of our approach and mitigate the inherent depth ambiguity problem, we incorporate additional priors such as a low-resolution depth map obtained from an RGB-D sensor.

#### 2.4.1.1 Shape-from-Shading in RGB-D Sensing.

The combination of SfS with RGB-D sensing has been the subject of several studies, such as [86, 164, 188, 239, 250]. All of these works aim to solve an optimization problem w.r.t. the parameter $\mathcal{X}$, which represents geometric aspects of the scene,

$$\min_{\mathcal{X}} \|I - L_{\mathrm{o}}(\mathcal{X})\|_2^2 + \lambda \|\mathcal{X} - \mathcal{X}_0\|_2^2 + E_{\mathrm{reg}}(\mathcal{X}), \tag{2.78}$$

where the first term in the objective function corresponds to the data term. The stated second term in the optimization problem is commonly referred to as the *depth/normal prior* term. It serves as a zeroth-order regularization term for the geometry, but rather than incorporating it into $E_{\mathrm{reg}}$, we explicitly present it here within the context of SfS in RGB-D sensing. It plays an important role in ensuring that the resulting geometry is consistent with the measured values from the depth sensor. The remaining quantities in the optimization objective remain unchanged as in the original SfS problem depicted in Equation (2.75). As mentioned earlier, incorporating knowledge from a depth sensor can help alleviate the ambiguities associated with SfS *e.g.*, as demonstrated in the Eikonal Equation (2.77). However, previous attempts to solve Equation (2.78), such as those presented in [86, 164, 188, 239, 250], assume that lighting and/or albedo are known or estimated in a preprocessing step, and that the color and depth images from the RGB-D sensor have the same resolution. In Section 3.1.2, we thoroughly discuss these works and emphasize their limitations, which will then be addressed in Chapter 5.

Figure 2.13: Exemplary PS images. The captures depict the same object as in Figure 2.11. Each image is already segmented and taken under a different illumination condition[20].

In light of the discourse on SfS, it is natural to question why a single image is used instead of multiple images. This inquiry can be addressed in the context of the PS problem, which we shall explore in the upcoming section.

### 2.4.2 Photometric Stereo

The problem of photometric stereo (PS) can be viewed as a natural extension of SfS. Consequently, many of the techniques developed for SfS, such as methods for recovering depth from normals, are naturally employed in conjunction with PS. The primary difference between SfS and PS lies in the number of input images. While SfS infers shape from a single image, Robert J. Woodham proposed in 1980 in his seminal work [238] to infer shape from multiple, differently illuminated images, an approach now known as *photometric stereo*. The original assumptions of PS proposed in [238] are that

- the object being captured is stationary and diffuse,
- the camera's relative position w.r.t. the object remains constant during image capture,
- different illumination conditions are used to capture each image, and
- the object is segmented and a foreground mask is provided.

Figure 2.13 shows some exemplary PS images. The objective is to acquire multiple images of an object under distinct illumination scenarios, as the reflectance and geometry of an object are presumed to remain unchanged under varying illumination conditions. This constraint provides sufficient information about the underlying scene assets to estimate the object's geometry and reflectance from a set of images, assuming the illumination is known. Several approaches exist that relax some of the aforementioned assumptions,

[20]Underlying geometry taken from `https://www.thingiverse.com/thing:897412/remixes`, accessed on 7th of July, 2023 at 4.01PM.

including those that utilize multiple camera viewpoints in a multi-view scenario [65, 113, 114, 125, 132, 168, 174, 247, 262], non-Lambertian approaches [43, 48, 135, 190, 208, 218, 259], light-source-enhanced cameras [8, 9, 27, 138, 157, 255], and dynamic scenes [78, 90, 101, 118, 227]. However, these approaches are beyond the scope of this thesis, which focuses on the following problem.

We may mathematically formalize the problem of PS by considering a set of $N$ images denoted by $\mathcal{I} = \{I_i\}_{i=1,\dots,N}$,

$$\min_{\mathcal{X}} \sum_{i=1}^{N} \|I_i - L_{\mathrm{o}}(\mathcal{X})\|_2^2 + E_{\mathrm{reg}}(\mathcal{X}), \tag{2.79}$$

where the parameter vector $\mathcal{X}$ contains information about the geometry such as normals or a depth map, similar to the SfS case. However, in the context of PS, we typically optimize over the albedo as well. It should be noted that the outgoing radiance $L_{\mathrm{o}}(\mathcal{X})$ is dependent on $i$ *i.e.*, the $i$-th incoming radiance, since it changes with varying illumination. Nonetheless, for the sake of clarity, we temporarily omit this dependence.
For a given configuration, the number of images $N$ required for disambiguating shape estimation varies. In the specific context of our investigation into directional (Section 2.3.3.1) and natural (Section 2.3.3.2) illumination, $N \geq 3$ and $N \geq 4$ images, respectively, are required[21]. In a matrix formulation, we can represent each quantity involved in the problem. Specifically, we can define the $N$ intensity images, each with $P$ pixels, as $\mathbf{I} \in \mathbb{R}^{N \times P}$, the surface normals as $\mathbf{N} \in \mathbb{R}^{3 \times P}$, the first-degree SH coefficients for each surface normal as $\mathbf{Y}_1 \in \mathbb{R}^{4 \times P}$, the albedo as $\boldsymbol{\rho} \in \mathbb{R}^{P}$ and construct the diagonal matrix $\mathbf{P} = \mathrm{diag}(\boldsymbol{\rho}) \in \mathbb{R}^{P \times P}$, and the lighting as $\mathbf{L}^{\mathrm{dir}} \in \mathbb{R}^{N \times 3}$ (for directional illumination) and $\mathbf{L}^{\mathrm{SH}} \in \mathbb{R}^{N \times 4}$ (for SH-based lighting). By defining $\mathbf{M}$ as a matrix indicating the scaled normals or SH basis functions with the albedo *i.e.*, $\mathbf{M}^{\mathrm{dir}} = \mathbf{NP}$ or $\mathbf{M}^{\mathrm{SH}} = \mathbf{Y}_1\mathbf{P}$, respectively, we can formulate the PS problem as a matrix multiplication,

$$\mathbf{I} = \begin{cases} \mathbf{L}^{\mathrm{dir}}\mathbf{M}^{\mathrm{dir}}, & \text{if directional light,} \\ \mathbf{L}^{\mathrm{SH}}\mathbf{M}^{\mathrm{SH}}, & \text{if SH light.} \end{cases} \tag{2.80}$$

To enhance conciseness, we use $\mathbf{M}$ to refer to both $\mathbf{M}^{\mathrm{SH}}$ and $\mathbf{M}^{\mathrm{dir}}$, and similarly for $\mathbf{L}$. It is worth noting that the element-wise clamping operation $\max(\cdot, 0)$ is neglected for the directional lighting case, which is a common simplification that introduces inaccuracies.

[21] In the case of natural light, the number of images required for shape estimation is generally greater than or equal to $(n+1)^2$, where $n$ represents the degree of SH used for modeling illumination, as discussed in Section 2.3.3.2. For the purposes of this argument, we restrict our attention to the first-degree of SH *i.e.*, $n = 1$.

To mitigate this, discrepancies between the data and the model can be described as sparse errors in a preprocessing step using [240]. The formulation in Equation (2.80) indicates that the image matrix $\mathbf{I}$ is, at most, of rank $3$ or $4$, as it can be expressed as a product of two low-rank matrices. Consequently, it is necessary to have at least $N \geq 3$ or $N \geq 4$ images to fully disambiguate the unknowns in the PS model, such that

$$\operatorname{rank}(\mathbf{I}) = \begin{cases} 3, & \text{if directional light,} \\ 4, & \text{if SH light.} \end{cases} \tag{2.81}$$

If the rank of the matrix is lower, then the linear system becomes underdetermined, and hence, no reasonable solution can be obtained. It is important to note that even if we have enough images, the rank of the matrices $\mathbf{L}$ or $\mathbf{M}$ may also be low. To avoid this situation, the illumination, shape, and reflectance should be complex enough. Inappropriate scenarios that can result in low rank matrices include planar lighting ($\operatorname{rank}\mathbf{L} = 2$), degenerate surfaces[22] ($\operatorname{rank}\mathbf{M} \leq 2$), vanishing albedo (constant $0$ *i.e.*, $\operatorname{rank}\mathbf{P} = 0$), inadequate pixel count, et cetera.

If the matrices have an appropriate rank, two scenarios exist for PS, namely, *calibrated photometric stereo (CPS)* when the illumination is known, and *uncalibrated photometric stereo (UPS)* when the illumination is unknown. Therefore, $\mathcal{X}$ in (2.79) can also comprise lighting information in the case of UPS. Although the absence of light calibration facilitates the overall capturing scenario in UPS, it can pose difficulties in solving for albedo, light, and shape. This can be exemplified by referencing (2.80). Any invertible matrix $\mathbf{A}$ can be used to preserve the resulting images as

$$\mathbf{I} = \mathbf{L}\mathbf{M} = \mathbf{L}\mathbf{A}\mathbf{A}^{-1}\mathbf{M}, \tag{2.82}$$

where we call $\mathbf{A}$ the *ambiguity matrix*. Depending on the lighting scenario, $\mathbf{A} \in \mathrm{GL}_3$ or $\mathbf{A} \in \mathrm{GL}_4$, where $\mathrm{GL}_n$ denotes the *general linear group* of degree $n$, which is the set of invertible $n \times n$ matrices. However, further constraints can be imposed on the shape matrices to alleviate the ambiguities illustrated in (2.82). Enforcing the integrability of surface normals under orthographic or perspective projection can be helpful. For example, when directional lighting is used and no integrability constraint is imposed, the ambiguity has nine degrees of freedom (dofs) [88]. However, if integrability is assumed under orthographic projection, the dofs reduce to three [252][23]. Interestingly, under

[22] Degenerate surfaces refer to simple geometries, such as a plane. We refer the interested reader to the work of Brahimi *et al.* [1].

[23] In the scenario where surface integrability is enforced under orthographic projection, the corresponding ambiguity matrix with three dofs is referred to as the generalized bas-reflief (GBR) ambiguity, as documented in [23].

perspective projection, enforcing integrability leads to a well-posed problem with no dofs [170]. When considering first-degree SH lighting, if no integrability is imposed on the surface normals, there are six dofs [20][24]. However, when enforcing integrability under orthographic or perspective projection, the dofs can be reduced to one and zero, respectively [1]. Enforcing integrability under perspective projection ensures the well-posedness of UPS in the directional and first-degree SH cases. However, the well- or ill-posedness under integrability remains an open research problem for higher-degree SH lighting as well as point light source illumination. In the absence of integrability enforcement, second-degree SH and point light source illumination yield nine and four dofs, respectively [20, 171]. While the specifics of these findings are beyond the scope of this thesis, the results have been included here for completeness, and a detailed discussion can be found in [1].
Thus far, we have expounded upon the assumptions and objectives of PS, the requisite number of images for a given setup, and provided a succinct overview of the inherent ambiguities that arise in the UPS scenario. Moving forward, we delve deeper into CPS under directional lighting, and UPS under natural illumination, as these are the fundamental building blocks for Chapter 6 and Chapter 7.

#### 2.4.2.1 Photometric Stereo under Directional Illumination.

In this section, we focus on the problem of recovering shape as a depth map $z$ in the context of CPS under directional light. To achieve this, we employ an image formation model of the form

$$I_i(\mathbf{p}) = \rho(\mathbf{p})\big\langle \mathbf{n}[z](\mathbf{p}), \mathbf{l}_i^{\text{dir}} \big\rangle, \tag{2.83}$$

which takes into account the $N$ given images $I_i$ and the associated light vectors $\mathbf{l}_i^{\text{dir}}$, where $i = 1, \ldots, N$. As stated earlier, the clamping operation involving $\max(\cdot, 0)$ has been omitted in this analysis, and instead, we employ a preprocessing step based on [240] to address arising inaccuracies. Let us substitute the expression for surface normals in terms of depth under orthographic projection (2.70) into Equation (2.83) and formulate it as a variational problem involving a set of non-linear PDEs,

$$\min_{\substack{z:\Omega\to\mathbb{R} \\ \rho:\Omega\to\mathbb{R}}} \sum_{i=1}^{N} \left\| I_i - \frac{\rho}{\sqrt{|\nabla z|^2 + 1}} \left\langle \begin{pmatrix} -\nabla z \\ 1 \end{pmatrix}, \mathbf{l}_i^{\text{dir}} \right\rangle \right\|_2^2 + \lambda \|z - z_0\|_2^2. \tag{2.84}$$

[24] The group of Lorentz transformations is employed to characterize the six dofs [61, 179].

The objective function includes a trade-off parameter $\lambda \geq 0$ and a linear least squares depth prior term. The dependence of the data term solely on the gradient of the depth, $\nabla z$, can give rise to a certain ambiguity in the minimization process. Specifically, since a constant can be added to the depth while retaining the same minimum, this ambiguity can be resolved by including a depth prior term. This term, governed by a positive parameter $\lambda$, enables fixing the constant by constraining the depth map to conform to a specific initialization value $z_0$ [144, 186]. Expressing (2.84) in terms of its depth map eliminates the necessity of a two-step approach involving estimating normals followed by integrating them, but introduces non-convexity as a result of the normalization factor $\sqrt{|\nabla z|^2 + 1}$. In addition, the optimization problem involves two variables, as the albedo $\rho$ is generally unknown. Therefore, in the absence of a good initialization of $\rho$ and $z$, convergence to poor local minima is possible.

To overcome this issue, we employ the technique of using *image ratios* or *photometric ratios*, which was first proposed in [54] and has since been successfully adapted to various PS settings [78, 133, 143, 144, 145, 146, 147, 148, 187, 211]. In particular, we follow the approach of [187], which employs image ratios in a variational setting for CPS under orthographic projection. Upon closer inspection of (2.83), it is possible to reformulate it such that it results in a set of PDEs that are not dependent on $\rho$ and are linear w.r.t. $\nabla z$, thus effectively circumventing the aforementioned issues. Specifically, by dividing the image by the shading, we can observe that this quantity is equal to the ratio of the albedo divided by the normalization factor. Per pixel, this ratio is constant across all images. We can express this relation between two different images $I_i$ and $I_j$, where $i \neq j$, as

$$\frac{I_i(\mathbf{p})}{\left\langle \begin{pmatrix} -\nabla z(\mathbf{p}) \\ 1 \end{pmatrix}, \mathbf{l}_i^{\text{dir}} \right\rangle} = \frac{\rho(\mathbf{p})}{\sqrt{|\nabla z(\mathbf{p})|^2 + 1}} = \frac{I_j(\mathbf{p})}{\left\langle \begin{pmatrix} -\nabla z(\mathbf{p}) \\ 1 \end{pmatrix}, \mathbf{l}_j^{\text{dir}} \right\rangle}. \tag{2.85}$$

It is worth noting that our approach assumes that the shading value is non-zero for every pixel in the image. This is a necessary assumption to ensure that the system being solved is non-degenerate. Specifically, degeneracy can occur in cases such as vanishing lighting where the light direction vector is a constant zero vector, or in the case of planar surfaces where all light vectors are perpendicular to the surface. By discarding the term in the middle of Equation (2.85), multiplying both sides by their respective denominator, and

performing further simplifications, we arrive at

$$\left\langle \underbrace{\begin{pmatrix} I_i(\mathbf{p})\mathbf{l}_{j,1}^{\text{dir}} - I_j(\mathbf{p})\mathbf{l}_{i,1}^{\text{dir}} \\ I_i(\mathbf{p})\mathbf{l}_{j,2}^{\text{dir}} - I_j(\mathbf{p})\mathbf{l}_{i,2}^{\text{dir}} \end{pmatrix}}_{\mathbf{a}_{ij}(\mathbf{p})}, \nabla z(\mathbf{p}) \right\rangle = \underbrace{I_i(\mathbf{p})\mathbf{l}_{j,3}^{\text{dir}} - I_j(\mathbf{p})\mathbf{l}_{i,3}^{\text{dir}}}_{b_{ij}(\mathbf{p})} \tag{2.86}$$

$$\langle \mathbf{a}_{ij}(\mathbf{p}), \nabla z(\mathbf{p}) \rangle = b_{ij}(\mathbf{p}), \tag{2.87}$$

which involves a vector field $\mathbf{a}_{ij} : \Omega \to \mathbb{R}^2$ and a scalar field $b_{ij} : \Omega \to \mathbb{R}$. The notation $\mathbf{l}_{i,k}^{\text{dir}}$ denotes the $k$-th element of the $i$-th directional light vector $\mathbf{l}_i^{\text{dir}} \in \mathbb{R}^3$, where $k \in \{1, 2, 3\}$. Doing this for every pair of images $(i, j)$ with $1 \leq i < j \leq N$, we can obtain $\binom{N}{2}$ equations. We define the set of all such pairs as $\mathcal{T} = \{(i, j) \,|\, 1 \leq i < j \leq N\}$. Substituting the $N$ data terms in (2.84) with the image-ratio-based formulation obtained in (2.87) yields an optimization problem of the following form:

$$\min_{z:\Omega\to\mathbb{R}} \sum_{(i,j)\in\mathcal{T}} \|\langle \mathbf{a}_{ij}, \nabla z \rangle - b_{ij}\|_2^2 + \lambda \|z - z_0\|_2^2. \tag{2.88}$$

Note that the number of equations in (2.84) increases linearly with $N$. However, in (2.88), we observe a quadratic behavior due to the cardinality of $\mathcal{T}$ being $|\mathcal{T}| = \binom{N}{2} = \frac{N!}{(N-2)!2!} = \frac{N(N-1)}{2}$. The problem stated in (2.88) is now a linear system w.r.t. $z$ and is independent of the albedo. Consequently, we can efficiently solve (2.88) using methods such as Cholesky decomposition or conjugate gradient iterations, as outlined in [187]. Through the joint solution of both PS and segmentation problems in Chapter 7, the presented image ratio model can help to effectively eliminate one of the fundamental assumptions underlying PS, namely the requirement of a foreground mask.

#### 2.4.2.2 Photometric Stereo under General Illumination.

This section deals with the problem of UPS under general illumination, which is commonly approximated using the SH model [21, 191], as discussed in Section 2.3.3.2. As in the previous section, to avoid the issue of non-integrable normal fields we take a differential approach, enforcing integrability by directly optimizing the underlying perspective depth map. The image formation model for the $i$-th image ($i = 1, \ldots, N$) is based on (2.63) and uses a second-degree SH approximation,

$$I_i(\mathbf{p}) = \rho(\mathbf{p}) \big\langle \mathbf{l}_i^{\text{SH}}, \mathbf{Y}_2(\mathbf{n}[z](\mathbf{p})) \big\rangle, \tag{2.89}$$

with the perspective surface normal, as shown in (2.73), but oriented towards the camera,

$$\mathbf{n}[z](\mathbf{p}) = \frac{\begin{pmatrix} \nabla_f z(\mathbf{p}) \\ -z - \langle \mathbf{p} - c, \nabla z(\mathbf{p}) \rangle \end{pmatrix}}{\sqrt{|\nabla_f z(\mathbf{p})|^2 + (z + \langle \mathbf{p} - c, \nabla z(\mathbf{p}) \rangle)^2}}. \tag{2.90}$$

The task of recovering the underlying depth map, albedo, and light vectors from a collection of $N$ images under the relation depicted by Equation (2.89), can be cast as a variational optimization problem of the form

$$\min_{\substack{z:\Omega\to\mathbb{R} \\ \rho:\Omega\to\mathbb{R} \\ \{\mathbf{l}_i^{\mathrm{SH}}\}_{i=1,\dots,N}}} \sum_{i=1}^{N} \left\| I_i - \rho \left\langle \mathbf{l}_i^{\mathrm{SH}}, \mathbf{Y}_2(\mathbf{n}[z]) \right\rangle \right\|_2^2. \tag{2.91}$$

Similar to (2.84), a depth prior term can be added to fix the multiplicative ambiguity w.r.t. the depth that arises under perspective projection. Despite the presence of this ambiguity w.r.t. depth, the solution of (2.91) poses a significant challenge due to non-convexity of the problem w.r.t. $z$. This non-convexity arises from the normalization factor in (2.90) and the second-degree SH basis functions, as illustrated in (2.56)–(2.60). The non-convexity of (2.91) necessitates a good depth initialization and a robust optimization scheme to obtain good solutions.
The stated challenging UPS problem (2.91) or similar variants have been addressed in numerous research works [7, 20, 154, 207]. In Section 3.2, we will conduct a comprehensive review of this relevant literature and analyze how the corresponding researchers have addressed the challenges associated with estimating the geometry of a scene with high accuracy in the presence of natural illumination. In summary, these methods employed multi-step approaches that are susceptible to error accumulation or relied on hardware-assisted shape initialization, which is not always feasible. The shortcomings of the existing methods will be overcome through a new paradigm proposed in Chapter 6, which incorporates a robust solver and a generic depth initialization approach.

## 2.5 Active Contour Segmentation

The objective of this section is to deliberate on a conventional strategy for the segmentation of objects, which relies on active contours as proposed by Chan and Vese [41]. This method has been effectively employed in the context of the research work presented in Chapter 7, to enable the concurrent estimation of geometry and its binary mask for PS.

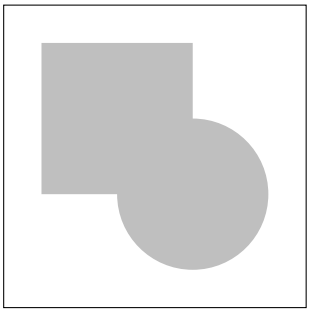


Figure 2.14: Example input image for Chan-Vese segmentation. This input image serves as an illustration for a specific scenario designed for Chan-Vese image segmentation, where the white region represents the background and the gray region depicts the foreground.

The main concept involves the partitioning of a given grayscale image $I : \Omega \subset \mathbb{R}^2 \to \mathbb{R}$ into two distinct regions, which are separated by a curve of minimal length. This curve is mathematically defined as the boundary of an open set $\tilde{\Omega} \subset \Omega$, denoted by $C = \partial\tilde{\Omega}$. Moreover, we define the interior and exterior regions w.r.t. the curve as $\mathrm{inside}(C) = \tilde{\Omega}$ and $\mathrm{outside}(C) = \Omega \setminus \mathrm{cl}\Big(\tilde{\Omega}\Big)$, respectively, where $\mathrm{cl}$ denotes the closure of a set. It should be noted that the image values are expected to remain nearly constant within and outside the curve, and are typically represented by the constants $\mu_1, \mu_2 \in \mathbb{R}$, respectively. The objective at hand is to determine the optimal curve $C$, as well as the associated constants $\mu_1$ and $\mu_2$, which can be formulated as the following optimization problem[25]

$$\min_{\substack{\mu_1,\mu_2\in\mathbb{R}\\ C}} \int_{\mathrm{inside}(C)} \mathcal{P}_{\mathrm{CV}}(\mu_1, I(\mathbf{p}))\,\mathrm{d}\mathbf{p} + \int_{\mathrm{outside}(C)} \mathcal{P}_{\mathrm{CV}}(\mu_2, I(\mathbf{p}))\,\mathrm{d}\mathbf{p} + \nu\,\mathrm{length}(C), \quad (2.92)$$

with $\mathcal{P}_{\mathrm{CV}}(\mu_i, I(\mathbf{p})) = |\mu_i - I(\mathbf{p})|^2$ for $i \in \{1, 2\}$, and the parameter $\nu \geq 0$ penalizing the length of the curve. Specifically, higher and lower values of $\nu$ correspond to shorter and longer curves, respectively. The proposed Chan-Vese segmentation method [41] is related to the classical Mumford-Shah model [156], which is commonly used for image segmentation tasks and results in piecewise smooth regions. To be more precise, imposing the constraint that the solution is piecewise constant leads to the well-known minimal partition problem, which can be regarded as a more general framework that encompasses the Chan-Vese model as a special case, where the partition is limited to only two distinct regions, see Figure 2.14.
To tackle this optimization problem, Chan and Vese [41] employ level set methods [166]

[25]The notation $\mathcal{P}_{\mathrm{CV}}$ is derived from the surnames of the original authors, namely Tony F. Chan and Luminita A. Vese, who proposed this method in [41]. In the literature, this model is commonly referred to as the Chan-Vese segmentation.

and utilize the reparametrization technique proposed in [258], which involves the use of the Heaviside step function[26]. Level set methods enable the representation of the curve $C \in \Omega$ as the zero level set of a Lipschitz function $\phi : \Omega \to \mathbb{R}$. This representation allows for the expression of $C$, $\mathrm{inside}(C)$, and $\mathrm{outside}(C)$ in terms of the level set function $\phi$ as

$$C = \{\mathbf{p} \in \Omega \,|\, \phi(\mathbf{p}) = 0\} \tag{2.93}$$

$$\mathrm{inside}(C) = \{\mathbf{p} \in \Omega \,|\, \phi(\mathbf{p}) > 0\} \tag{2.94}$$

$$\mathrm{outside}(C) = \{\mathbf{p} \in \Omega \,|\, \phi(\mathbf{p}) < 0\}. \tag{2.95}$$

By utilizing (2.93) – (2.95), it is possible to restate the initial problem (2.92) in terms of the level set function $\phi$,

$$\min_{\substack{\mu_1,\mu_2\in\mathbb{R}\\ \phi:\Omega\to\mathbb{R}}} \int_{\phi>0} \mathcal{P}_{\mathrm{CV}}(\mu_1, I(\mathbf{p}))\,\mathrm{d}\mathbf{p} + \int_{\phi<0} \mathcal{P}_{\mathrm{CV}}(\mu_2, I(\mathbf{p}))\,\mathrm{d}\mathbf{p} + \nu\,\mathrm{length}(\phi = 0). \tag{2.96}$$

Despite the reformulation of the initial problem in terms of the level set function $\phi$ as given in (2.96), finding a solution remains challenging due to the optimization variable dependence on the integration domain. To address this difficulty, we can leverage the Heaviside step function

$$H : \mathbb{R} \to \{0, 1\},\ H(x) = \begin{cases} 1, & \text{if } x \geq 0, \\ 0, & \text{if } x < 0. \end{cases} \tag{2.97}$$

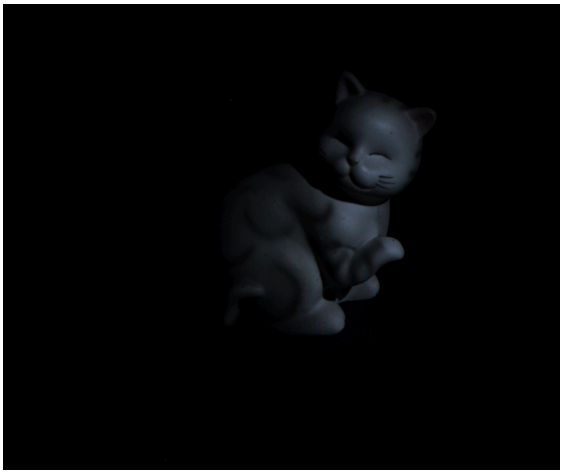
By smartly multiplying $H$ with the integrands from (2.96), we can integrate over the entire image domain $\Omega$

$$\min_{\substack{\mu_1,\mu_2\in\mathbb{R}\\ \phi:\Omega\to\mathbb{R}}} \int_{\Omega} H(\phi(\mathbf{p}))\mathcal{P}_{\mathrm{CV}}(\mu_1, I(\mathbf{p})) + (1 - H(\phi(\mathbf{p})))\mathcal{P}_{\mathrm{CV}}(\mu_2, I(\mathbf{p})) + \nu|\nabla H(\phi(\mathbf{p}))|\,\mathrm{d}\mathbf{p}, \tag{2.98}$$

where we used $\mathrm{length}(C) = \mathrm{length}(\phi = 0) = \int_\Omega |\nabla H(\phi(\mathbf{p}))|\,\mathrm{d}\mathbf{p}$ to express the length of the curve $C$ in terms of the level set function $\phi$.

The optimization problem in (2.98) is non-convex, and the solution heavily depends on the initialization, thus leading to the possibility of getting stuck in local minima. Furthermore, this approach tends to work well only if the image can be described using two values, as in the resulting values of $\mu_i$, $i \in \{1, 2\}$, which closely correspond to the av-

[26] The Heaviside step function is named after the British mathematician and physicist Oliver Heaviside (1850–1925).



(a) Image to be segmented[27].

(b) Segmented object of the image in (a).

Figure 2.15: Illustration of fading foreground and background. In (a), the fading foreground on the left side of the cat presents challenges in distinguishing between the background and foreground in that region. For reference, the segmented object is visualized in (b), revealing that the color of the region, which resembles the background, actually belongs to the foreground. This characteristic poses difficulties for active contour methods to accurately perform object segmentation.

erage value of $I$ inside and outside the curve. However, this model may fail to perform effectively in the presence of fading or discontinuous boundaries, where no clear distinction can be made between the foreground and background, as illustrated in Figure 2.15. In the context of solving PS and masking simultaneously, we present in Chapter 7 an alternative approach based on a customized Chan-Vese model that effectively addresses the problem above. While this alternative approach entails the utilization of a more sophisticated cost function, it removes one of the fundamental assumptions of PS, namely the masking process.

[27]Image taken from [208].

# Chapter 3

# Related Work

This chapter offers a comprehensive overview of the relevant prior work that serves as the foundation for the main publications presented in Part II. To begin, we examine the related work on depth super-resolution (SR) and Shape-from-Shading (SfS) techniques. Next, we present the most advanced methods for uncalibrated photometric stereo (UPS) under natural illumination. Furthermore, we discuss the generation of object masks in a photometric stereo (PS) setting and emphasize the related limitations. Lastly, we present the current standard for material estimation in large-scale environments.

## 3.1 Depth Super-Resolution and Shape-from-Shading

This section offers a comprehensive review of the current state-of-the-art in depth SR and SfS. Specifically, we focus on the intersection of both topics, as both problems are addressed simultaneously in Chapter 5. We begin by discussing image-guided depth SR, which leverages corresponding RGB images to aid in depth SR. Next, we present the prior research that deals with SfS with a depth prior, for instance, obtained from an RGB-D sensor.

### 3.1.1 Image-Guided Depth Super-Resolution

The problem of image-guided depth SR has become increasingly relevant with the availability of low-cost RGB-D cameras. These cameras offer aligned pairs of RGB and depth images, which can be exploited to solve the depth SR problem by leveraging the high-resolution information provided by the RGB image. A common assumption in this case is to enforce alignment of depth edges with edges in the corresponding RGB image, such that the depth image is smooth in regions where the RGB image is smooth, and discontinuous in regions where the RGB image is discontinuous [58, 70, 172, 173, 246].

**Diebel and Thrun [58].** In Diebel and Thrun [58], the problem of image-guided depth SR was formulated using Markov random fields (MRFs). The formulation includes terms similar to (2.9) and (2.10), with an additional weighting factor based on intensity variations in the corresponding grayscale image.
As a result, regions in the depth map with little intensity variation in the grayscale image are smoothed more, while regions with large intensity variation are subject to less smoothing.

**Yang *et al.* [246].** An approach based on cost volumes was proposed in [246]. In this method, an initially manually upsampled depth map is used to construct a cost volume based on a truncated quadratic cost function. Each slice of the cost volume is then filtered using a bilateral filter that is dependent on the corresponding color image. The bilateral filter results in smooth and discontinuous depth maps in regions where the image is smooth and discontinuous, respectively. The best cost is then extracted from the filtered cost volume and refined using a sub-pixel refinement to obtain a new depth map. This process is repeated multiple times, with each new depth map being used to construct the subsequent cost volume.
Although the results of [246] outperform the MRF approach proposed in [58], oversmoothing is still noticeable due to the underlying bilateral filter being applied.

**Park *et al.* [172, 173].** The issue of oversmoothing in image-guided depth SR was addressed in [173] and its ensuing work [172] by introducing a more sophisticated objective function. While the data term closely follows (2.9), two additional prior terms are introduced. The first is a regularization term based on non-local means (NLM) [36] weighted with an anisotropic structural-aware filter [46]. The anisotropic filter relies on image gradients, while the NLM filter compares pixels in a neighborhood, including non-first-order neighbors. The second regularization term is a weighted version of (2.10). The weight is based on confidence and reflects the spatial coherence of first-order neighbors. The confidence weighting is determined by color similarities, superpixel-based segmentation [12, 13], edge saliency [26], and a guided bicubic interpolated depth map.
Although [173] yields sharper results than [58, 246], the reliance on color image weights can cause erroneous propagation of color information to the depth image, leading to inaccurate estimation of depth discontinuities.

**Ferstl *et al.* [70].** A more mathematically-oriented approach, as opposed to the "weight-engineering" approach of [173], was presented in [70]. The data term is similar to (2.9), while the regularization term is based on a generalization of the total variation

(TV) prior, known as total generalized variation (TGV) [33]. TGV enables the reconstruction of piecewise polynomial functions[1]. In [70], TGV is used to fit piecewise affine functions to a coarse depth map. Furthermore, the TGV regularization is weighted with an anisotropic diffusion tensor [192, 236], which assumes that color discontinuities coincide with depth discontinuities.

**Summary.** The assumption of depth discontinuity coinciding with color discontinuity is a common thread among the works discussed in this section, namely [58, 70, 172, 173, 246]. While this assumption may hold in some cases, it is not always true since variations in the RGB image may not necessarily correspond to variations in the depth image, as in the case of a simple drawing on a sheet of paper. Moreover, this assumption only utilizes sparse information from the RGB image, whereas the entire RGB image contains valuable geometric information. To address this limitation, some works have leveraged physically-based approaches to densely relate the RGB image to the scene's assets, such as shape, illumination, and material (Section 2.3). These approaches, including [7, 8, 137], are primarily focused on the problem of PS and require multiple images to solve depth SR. However, in the upcoming section, we will explore the case of utilizing only a solitary RGB-D pair and combining it with the technique of SfS to retrieve a more detailed depth map.

### 3.1.2 Shape-from-Shading in RGB-D Sensing

In the following, we will examine related research that addresses the challenge of solving the SfS problem while assuming the availability of a rough geometry estimate. As previously discussed in Section 2.4.1, the SfS problem is highly under-determined, since an image can be explained by any geometry if the albedo and illumination are sufficiently complex. Therefore, the approach of incorporating RGB-D data with SfS is to guide shape optimization towards the correct direction using a sound geometry prior. Despite the potential for high-quality 3D reconstruction from a single RGB-D image pair, this problem scenario has not been extensively studied, with only a few works addressing it [86, 164, 188, 239, 250].

**Yu *et al.* [250].** The pioneering work that applied SfS with RGB-D data was [250]. Their method involves a multi-stage process where first the RGB image and a coarse normal map based on the depth map are used to iteratively estimate albedo and spherical harmonics (SH) illumination. They leverage the fact that pixels with similar normal

[1]The $k$-th order TGV favors piecewise polynomials of order $k-1$. Thus, TGV is a generalization of TV, which results in piecewise constant functions *i.e.*, first-order TGV.

directions have the same shading under consistent natural illumination, and the differences in their pixel intensities are due to differences in their respective albedos. To achieve this objective, the authors employed a clustering method to group similar image intensities and then constructed a graph, wherein each cluster represented a node and two nodes were connected if they shared multiple normal directions. This graph was then utilized to identify clusters of albedos *with uniform color*, resulting in a piecewise constant reflectance assumption. The obtained albedos are further utilized to estimate SH coefficients, which can be used to refine the albedos again. This iterative process is repeated for $3$–$5$ times to achieve piecewise constant albedo and SH coefficient estimates. Note that the used normals are based on the RGB-D sensor's depth map, which is subject to noise and quantization artifacts (Section 2.2). Concurrently, they perform patch-based depth map repairing to estimate missing regions. These regions are first inpainted by estimating depth gradients based on RGB data, RGB smoothness, depth gradient data, and depth gradient smoothness. Afterwards, the depth gradients are used to recover absolute depth via Poisson integration. Finally, the estimated albedo, SH lighting, and repaired depth map of the two previous steps are used to estimate a normal map. The method minimizes a loss function similar to (2.75), where the data term is based on a second-degree SH approximation (2.62). Their image formation model w.r.t. surface normals under precomputed albedo and lighting is $L_\mathrm{o}^{[250]}(\mathbf{n}) = \rho\langle \mathbf{l}^\mathrm{SH}, \mathbf{Y}_2(\mathbf{n})\rangle$. The method infers three regularization/prior terms on $\mathbf{n}$: a normal prior as in (2.78), and two additional terms ensuring smoothness and unit length.
In the proposed overall approach, the albedo and lighting estimation does not benefit from the refined normals, as only the coarse sensor depth is utilized in this step. Moreover, the surface refinement step is limited to the estimation of surface normals, which may result in a non-integrable normal field. Furthermore, the employed smoothness term during normal estimation leads to overly smooth normals and may cause a loss of fine details. The authors have used data from the Kinect V1 throughout the paper, which captures RGB and depth at the same VGA resolution. However, as demonstrated in Table 2.1, all other depth sensors capture RGB and depth data of different resolutions. Therefore, it is not clear how [250] would perform if the resolutions were to differ.

**Han *et al.* [86].** The subsequent work [86] proposes a method for estimating high-quality normals from a single RGB-D image pair using a per-pixel $\alpha$ weighted second-degree SH model, assuming a constant albedo $\rho = 1$, $L_\mathrm{o}^{[86]}(\mathbf{n}) = \alpha\langle \mathbf{l}^\mathrm{SH}, \mathbf{Y}_2(\mathbf{n})\rangle$. The assumption of a constant albedo is typically realized by assuming a white albedo $\rho = 1$, and in the case where the object of interest is colored, colored lighting is used. Initially, a global SH illumination $\mathbf{l}^\mathrm{SH} \in \mathbb{R}^9$ is estimated using the constant albedo assumption and an initial normal map from the RGB-D sensor, which is solved in a least squares

manner with $\alpha = 1$. Subsequently, local lighting $\alpha : \Omega \to \mathbb{R}$ is estimated, where an adaptive smoothness based on the RGB image and a squared $L_2$-loss of the Laplacian of $\alpha$ are introduced to avoid overfitting to the RGB image and to avoid abrupt changes w.r.t. pixel positions. In the final optimization step, the normal map is refined by leveraging both the global and local light estimations. This process entails minimizing a least squares photometric loss that measures the residual between the estimated shading $L_{\mathrm{o}}^{[86]}(\mathbf{n})$ and the input intensity image. To further ensure the quality of the estimated normal map, two regularization/prior losses are incorporated: a normal prior as in (2.78) and an integrability constraint. The latter is enforced by minimizing the curl of the normals to ensure that the resulting normal field can be integrated to a surface, which is advantageous over [250]. Additionally, a unit length constraint is inferred by optimizing the normals in the $pq$-space [94], where a surface normal can be calculated as $\mathbf{n}(p, q) = \mathrm{normalize}(p, q, -1)^\top$ under orthographic projection[2].
Although the proposed approach provides highly detailed results, it has limitations that need to be taken into account. First, the constant albedo assumption must be ensured beforehand, which restricts the applicability of the method to a certain category of objects. Second, the normal parameterization employs orthographic projection while the depth sensor's depth is perspective, which makes it challenging to compare raw depth versus estimated depth without pre- and/or postprocessing to bring both surfaces to the same parameterization. As in [250], the global and local lighting estimation is solely based on the initial sensor's depth. Although the proposed algorithm could be deployed in an iterative scheme to refine the illumination using the refined normals and, in turn, further improve the estimated normals, this possibility is not discussed in the paper. Finally, it is worth noting that the approach assumes equal resolution between RGB and depth, which is another limitation similar to [250].

**Wu *et al.* [239].** The article by Wu *et al.* [239] presents an approach that achieves real-time depth refinement of RGB-D sensor data based on SfS using the graphics processing unit (GPU). Their method is designed for both single RGB-D pair and multi-view scenarios, where they incorporate temporal consistency terms. However, since our focus is not on temporal information, we will exclude this from our discussion. This exclusion does not affect the overall algorithm presented here, and the advantages and drawbacks mentioned here still apply. As with previous works, this method also assumes a second-degree SH illumination model. However, unlike the methods proposed in [86, 250], the proposed approach expresses the outgoing radiance in terms of perspective depth rather than in terms of normals. Specifically, their method computes the output radiance as

[2]If the normal field is integrable, $p$ and $q$ can be identified as the directional derivatives of a depth map, as described in Equation (2.70).

$L_{\mathrm{o}}^{[239]}(z) = \rho\langle \mathbf{l}^{\mathrm{SH}}, \mathbf{Y}_2(\mathbf{n}[z])\rangle$. The estimation of lighting is performed by solving a linear least squares problem. However, due to the lack of any assumptions on the albedo, it becomes challenging to estimate global illumination using a given depth map and a corresponding RGB image. As a result, the authors opt to assume a constant albedo of $\rho = 1$ during the lighting estimation step. Upon prediction of lighting, a more reliable estimate of the albedo is computed by dividing the intensities of the image by the shading induced by the predicted SH model. This computation is carried out point-wise, as no additional regularization is enforced. The subsequent depth estimation step comprises of a data term and two regularization terms. Interestingly, instead of imposing an absolute photoconsistency term, a shading gradient term is employed. This term penalizes the difference between the gradients of the rendered image and the input image. The justification for using the shading gradient term is its ability to handle deviations between the shading model and the real-world data more robustly. In contrast to previous works [86, 250], the authors propose an optimization scheme on the perspective depth values as the outgoing radiance depends on $z$, $L_{\mathrm{o}}^{[239]}(z)$. This approach has certain advantages in terms of integrability, as discussed in Section 2.3.4. To enhance robustness, a smoothness constraint penalizing the Laplacian of the depth map is proposed, along with a depth prior similar to that in [86, 250], as shown in (2.78). The authors then apply a Gauss-Newton scheme to solve for the depth. These steps, including lighting, albedo, and normal estimation, are implemented on the GPU with considerable engineering effort to make the proposed approach real-time capable.
In contrast to the approach of [86], the proposed method does not make the assumption of a constant albedo. Instead, a trivial albedo is only assumed during the lighting estimation step. The point-wise computation of the albedo, however, leads to overfitting to the data term and may absorb geometric information that cannot be recovered during the step that follows for estimating depth. Once the lighting and albedo have been estimated, the depth is updated as a last step. However, as with the methods in [86, 250], the albedo and lighting estimation steps are not being refined using the high-quality depth obtained in the previous step, as no iterative scheme across lighting, albedo and depth is carried out. Furthermore, the assumption of equal resolution between RGB and depth data persists [86, 250].

**Or-El *et al.* [164].** The research presented in [164] is another example of a real-time framework for generating high-quality depth maps from RGB-D sensors. Their proposed image formation model is based on first-degree SH lighting, with an additional pixel-wise shift $\beta : \Omega \to \mathbb{R}$ that takes into account specular highlights and local illumination changes. This model is motivated by the work of Grosse *et al.* [81] and then extended to SH illumination. Specifically, the outgoing radiance is expressed as

$L_{\mathrm{o}}^{[164]}(z) = \rho\langle \mathbf{l}^{\mathrm{SH}}, \mathbf{Y}_1(\mathbf{n}[z])\rangle + \beta$. Similar to the other methods discussed, the estimation of the lighting vector $\mathbf{l}^{\mathrm{SH}} \in \mathbb{R}^4$ is the first step in the proposed approach. This is achieved through solving a linear least squares problem using the depth map provided by the sensor, while assuming a constant albedo of $\rho = 1$ and no pixel-wise shift *i.e.*, $\beta = 0$. The next step after lighting estimation is albedo estimation, wherein a regularization term is incorporated to assume local smoothness of the albedo w.r.t. image and depth intensities. Specifically, if depth or image values are spatially similar, the resulting albedo should be smooth. This adaptive smoothness on the albedo is inspired by the local lighting illumination approach proposed in [86]. It is noteworthy that the assumption of $\beta = 0$ is still being upheld during this process. Prior to optimizing for shape, the parameter $\beta$ is estimated using a similar approach as that of the albedo estimation, accompanied by an additional regularization term that minimizes the squared $L_2$-loss of $\beta$, promoting small values for $\beta$. Orthographic surface refinement is conducted directly over depth values, circumventing the potential non-integrability issue present in approaches that rely on normal vectors, such as those proposed in [239]. The optimization procedure leverages a data term similar to those employed in [86, 250], which penalizes the difference between measured intensity and the rendering obtained from the estimated depth, given the fixed albedo and lighting parameters. To regularize the depth map, the common depth prior term, used in [86, 239, 250], is incorporated, and smoothness is inferred by minimizing the squared $L_2$-loss of the Laplacian of the depth map, similarly to [239]. Since the resulting optimization problem w.r.t. $z$ is non-convex, a lagged optimization strategy is employed. In this strategy, all non-convexities are fixed to their values from the previous iteration. The only non-convex part of the optimization problem is the normalization factor when computing normals from depth. Hence, the normalization factor from the last iteration is used, resulting in a linear least squares problem that can be efficiently solved.

This approach assumes an orthographic depth map and therefore requires preprocessing of the perspective depth map obtained from the RGB-D sensor, which is a similar issue to the one discussed in [86]. The lagged optimization approach employed in this method only infers the local surface orientation from its linear part, with the non-linear part fixed to the last iteration. However, this approach may miss useful information that could be contained in the non-linear parts. Furthermore, as with the previous methods discussed in [86, 239, 250], the lighting and albedo estimation steps do not benefit from the refined depth since no iterative scheme is applied, *and* the assumption of equal resolution between the RGB image and depth map remains.

**Quéau *et al.* [188].** A versatile and robust variational scheme is proposed in [188], which assumes that *both lighting and a constant albedo are known* upfront. The authors

propose to solve for the SH illumination case of first and second degree, using both orthographic and perspective projection of the depth. Specifically, the proposed method computes the radiance as $L_{\mathrm{o}}^{[188]}(z) = \rho\langle \mathbf{l}^{\mathrm{SH}}, \mathbf{Y}_n(\mathbf{n}[z])\rangle$, where $n \in \{1, 2\}$. The projection depends on the input data, making the approach applicable to a wider range of applications. In a similar fashion to [164, 239], the optimization in this method is performed over the depth map to ensure integrability. To disambiguate depth estimation and to handle noise, two regularization terms are employed. Similarly to previous works [86, 164, 239, 250], a common depth prior is used. However, instead of a smoothness term based solely on the gradient or Laplacian of the depth, a minimal surface regularization term is utilized [79]. This term, as seen in Equation (2.13), ensures that fronto-parallel solutions or similar problematic solutions are avoided, as discussed in Section 2.2.3.
Although this approach appears to be effective, its applicability is inherently limited as it assumes constant and known albedo and lighting, which restricts its utility in a narrower range of scenarios. Furthermore, similar to other methods discussed in this context [86, 164, 239, 250], this method also enforces the constraint that RGB and depth data must have the same resolution.

**Summary.** To provide a summary, all the approaches discussed here [86, 164, 188, 239, 250] suffer from two significant drawbacks. First, albedo and light are either assumed to be constant or given, or estimated only once at the beginning and then remain unaltered throughout the algorithm. It is worth noting that a more precise reconstruction of geometry can further refine the estimate of albedo and lighting, and vice versa. Therefore, it is desirable to update the values of albedo, lighting, and shape iteratively so that each variable can benefit from a more refined version of the others. However, none of the approaches discussed here employ such a framework, which limits the accuracy of the results. Second, all approaches assume that the resolution of the depth map and the RGB image are identical. However, an RGB-D sensor's depth map often has holes that must be inpainted before linearly interpolating the depth map to match the RGB's resolution. Thus, a range of preprocessing steps are necessary to increase the resolution of depth to that of the RGB image, which can alter the depth map's values in an unintended way. Therefore, in such cases, it is common to downsample the RGB image to the depth map's resolution to enable using the sensor's raw depth values, making the resolution of the depth map a limiting factor for accuracy and detail.

Considering the discussion on SfS with RGB-D data and image-guided depth SR in Section 3.1.1, it is proposed in Chapter 5 to address these limitations. This approach results in high-resolution and quality depth maps along with albedo and lighting estimates of unprecedented detail. Further information on how this is accomplished can be found in Chapter 5.

## 3.2 Uncalibrated Photometric Stereo under General Illumination

We will now review the literature related to the inversion of UPS under natural illumination problems, such as (2.89), and explore the methods used to overcome the inherent challenges caused by its ambiguity, integrability, and non-convexity. The works that address this complex inverse problem include [7, 20, 154, 207]. These approaches typically involve either a matrix formulation based on surface normals (as shown in (2.80)) and a multi-stage pipeline or a variational approach based on depth, similar to (2.91).

**Basri *et al.* [20].** The pioneering work of [20] addressed the problem of UPS under both first- and second-degree SH illumination. They adopted a linear algebra perspective by formulating the problem as a matrix factorization, $\mathbf{I} = \mathbf{L}^{\text{SH}}\mathbf{M}^{\text{SH}}$, as seen in (2.80). For the first-degree case, they proposed an algorithm based on the singular value decomposition (SVD) of $\mathbf{I}$, leveraging the low-rank property of $\mathbf{I}$ as in (2.81). However, the estimates for $\mathbf{L}^{\text{SH}}$ and $\mathbf{M}^{\text{SH}}$ obtained through this approach often lack two essential properties: the hypercone constraint[3] and the integrability of the normal field. The hypercone constraint ensures the separability of the shape matrix into albedo and SH functions, which in turn enables the recovery of surface normals, $\mathbf{M}^{\text{SH}} = \mathbf{Y}_1\mathbf{P}$. While the hypercone constraint can only be satisfied in the presence of little noise, the integrability constraint is crucial for the existence of an underlying surface (Section 2.3.4), and its satisfaction relies on human interaction to guess hand-chosen surface normals. Similarly, for second-degree SH, no constraint like the hypercone constraint exists to help infer the underlying shape up to integrability. In this case, a general-purpose optimization framework is employed to solve for a linear transformation that extracts the underlying shape from the SVD solution. As there is no hypercone constraint, one can only identify the (possibly non-integrable) normals up to a linear transformation. The integrability issue is resolved manually through a user-interactive postprocessing step, similarly to the first-degree case.

**Shi *et al.* [207].** The research conducted by [207] employs the use of PS under natural lighting conditions to a set of internet images to reconstruct objects from a variety of touristic sites, such as Kōtoku-in in Japan, Motherland Calls in Russia, the Taj Mahal in India, the Statue of Liberty, and Mount Rushmore in the USA. The objective of the

[3] It is called the hypercone constraint because it refers to a hypercone in 4D that is represented by the equation $x^2 + y^2 + z^2 = w^2$, which the first-degree SH satisfy with $w = 1$. The constants in (2.56)–(2.60) are absorbed into the lighting vector and can be neglected. When the SH are multiplied by the albedo, it simply scales the hypercone, resulting in $w = \rho$.

study is to resolve the issues associated with recovering shape information from images captured under natural lighting conditions using a shape prior that provides a good initialization. This approach eliminates the need for manually selecting normal vectors in images, which is a requirement in the previous work [20]. Initially, a set of internet images of the object of interest is downloaded, and a combination of sparse reconstruction techniques such as structure from motion (SfM) [212] and multi view stereo (MVS) [73] algorithms are applied. The resulting sparse point cloud is then utilized to create a watertight depth prior using Poisson surface reconstruction [115]. The calibrated images from SfM and MVS, along with the depth prior, are used to register the internet images and warp them to a manually selected reference view. The depth prior is also used to calculate a normal prior, which can assist in resolving the ambiguities associated with UPS. Specifically, a similar strategy to [20] based on SVD is employed, but with the addition of normal prior information from the SfM and MVS frameworks to disambiguate the linear transformation. Once the normal field is estimated, it is integrated into an orthographic depth map. Additionally, the depth prior is used as anchor points for surface recovery.
As already mentioned, [207] nicely estimates the linear ambiguity of UPS using a normal prior, as compared to [20]. However, the proposed technique requires significant preprocessing steps (SfM, MVS, Poisson integration, and registration) to retrieve an initial shape of the scene. Furthermore, solving the UPS problem involves two steps: estimating the normal map and then integrating it into a depth map.

**Peng *et al.* [7].** In their work, [7] leverage modern hardware to yield a robust depth prior by employing data from RGB-D sensors that provide synchronized RGB and depth information (Section 2.2). They utilize first-degree SH and present UPS as a variational problem in the form of (2.91) while including an additional SR depth prior term (2.9), resulting in a fully data-driven model. Furthermore, they directly optimize over depth to address the integrability issue, assuming perspective projection. Their variational approach allows for incorporating RGB data, a feat not easily achieved with SVD methods [20, 207][4]. The resulting variational framework can be solved in an alternating manner over albedo, light, and depth using a fixed point approach. This approach adopts a similar technique as in [164] in which the non-linear parts are trailing one iteration behind.
While this approach is transparent and optimizes over depth assuming perspective projection, it only utilizes first-degree SH, resulting in a lack of accuracy in the light approximation. Moreover, this approach heavily depends on a well-initialized shape due to the use of RGB-D sensor data, which may be difficult to obtain without such hardware, such

[4]In conventional SVD-based approaches, the RGB images are commonly converted to grayscale images, leading to the loss of useful information, as pointed out by Quéau *et al.* [187].

as in the case of [207]. Additionally, due to the least squares loss function, this approach is susceptible to outliers, which often manifest as cast shadows or specular highlights for non-diffuse surfaces.

**Mo *et al.* [154].** The first work to solve UPS under natural illumination without a sparse set of manually hand-chosen normals [20] or an automatically generated depth prior [7, 207] was proposed by Mo *et al.* [154]. The method follows a multi-step procedure, relying on the proposed *equivalent directional lighting* approximation. The authors assume that for small image patches, the underlying normals face roughly the same direction. Therefore, for each patch, the normal direction is approximated with a constant, and within each patch, the integration (2.38) is carried out over the same visible hemisphere oriented along the surface normal. This is claimed to be equivalent to directional lighting within a patch, where the light vector is the mean lighting over the visible hemisphere, as shown in (1) and (2) of Chapter 6.3. This leads to a patch-wise UPS problem under directional lighting, which is then solved separately per patch via an SVD, similar to the approach in [88]. For each patch, the resulting normal estimates are then ambiguous up to rotation. A rotation within a patch rotates all normals equally, hence the relative angle between two normals within a patch remains unaffected. A matrix of all pairwise angles across the patches is generated by leveraging 1) the rotation invariance of two relative angles within a patch and 2) the overlap between patches. From the given pairwise angles, they then follow [136] to compute a normal field up to a global rotation. As in [136], the global rotation can be fixed via enforcing integrability [23]. Finally, only a concave/convex ambiguity remains, which is resolved manually.
While the method by Mo *et al.* [154] can estimate normal maps without the need to explicitly estimate lighting and albedo, it suffers from accumulated errors in each step of the algorithm. This error accumulation can cause the resulting normal field to be poorly integrable and, therefore, far from the ground truth surface. Additionally, in each patch in the first step of the algorithm, the authors not only assume a constant normal field but also constant albedo. This assumption is crucial to estimate each patch's normal field up to rotation [88]. The assumption of a constant albedo and set of constant normals per patch also causes the algorithm to fail if more complex surfaces or albedos are present.

**Summary.** In conclusion, the current state-of-the-art methods are limited in their applicability to real-world UPS scenarios under natural illumination due to various assumptions. The existing methods rely on SVD solutions based on grayscale images followed by a second, independent step to ensure integrability [20, 154, 207]. Some methods use a transparent differential variational scheme to directly estimate depth, which eliminates the need for a two-step process [7]. However, such methods require a reliable

depth initialization from an RGB-D sensor, which may not always be feasible. Thus, it is essential to develop an approach based solely on RGB data that guarantees an integrable surface and is robust to outliers such as cast shadows or reflections. Despite this challenge, no existing UPS approach has addressed this issue, except in the simpler, directional lighting case [190]. In Chapter 6, we propose a method that satisfies these requirements through a generic minimal surface depth initialization with a single tuning parameter [167]. Further details on how this is achieved can be found in Chapter 6.

## 3.3 Masking for Photometric Stereo Approaches

As posited by Horn [92] and Woodham [238], photometric techniques such as SfS and PS rely on the availability of a foreground/background mask to separate the object to be reconstructed from the rest of the scene. The rationale behind this requirement is that the underlying depth map should exhibit smoothness without any discontinuities, which would be violated if the object and the background were "merged" together in the absence of segmentation. Thus, providing an accurate mask is pivotal to obtain high-quality geometry estimates. However, the masking process can be arduous, often performed manually or as a preprocessing step using various techniques, such as imaging software [220], statistical shape segmentation [161], or more recently, deep neural networks [213, 214]. In this section, we will delve into these techniques, highlight their respective merits and limitations, and conclude by elucidating the challenges that an algorithm should tackle to obviate the need for masking altogether.

**GIMP [220].** In many cases, the manual masking process is performed using image editing software such as GIMP [220]. Despite its accuracy in enabling the user to make pixel-level decisions, this approach is known to be highly time-consuming, and often requires multiple attempts before a satisfactory mask is obtained.

**Nieuwenhuis and Cremers [161].** GIMP also provides a semi-automated approach called the "Foreground Select Tool" that uses spatially varying color distributions for interactive segmentation [161]. However, this method still relies on user interaction to provide an initial guess of the color distribution for foreground and background, which can be difficult to estimate. In the case of PS, objects are typically placed in a dark room and illuminated with a single light source [208], leading to strong intensity variations that can cause shadows to blend seamlessly with the background. As a result, it is often difficult to separate the color distributions of the foreground and background, making such approaches unsuitable for PS data. If PS images are not taken in a dark

room, one would need to position the objects such that they are clearly distinct from the background in terms of color, highlighting the importance of considering the masking process during the acquisition setup.

**Sofiiuk [213] *et al.* and Song *et al.* [214].** In recent years, deep neural networks have been employed to perform image segmentation. However, only a few of these networks allow the user to iteratively refine the segmentation process, if the results are not accurate enough [213, 214]. Despite the high priority of accuracy in the context of PS, neural networks still struggle to achieve satisfactory results [153]. Moreover, segmentation networks are trained on images from daily life that are not necessarily related to PS, which makes them vulnerable to the same issue as [161] when dealing with shadows that blend with the background.

**Summary.** Overall, we can identify two limitations of the masking process in PS. First, it necessitates a manual or only partially automated (iterative) preprocessing step. Second, all available automated masking tools suffer from inaccuracies and non-robustness w.r.t. large image variations where the foreground blends with the background. These limitations are unsurprising, given that all approaches lack knowledge about the 3D scene and rely solely on image intensities. To overcome these challenges, it is preferable to avoid masking as a preprocessing step and instead mask the image on the fly while solving PS. Moreover, the accuracy of the mask and the final recovered shape should not be compromised, *and* the approach should remain robust w.r.t. the large image variations in PS data. To achieve this, we incorporate 3D knowledge of the scene into the problem, which leads to superior segmentation accuracy, while the reconstruction quality can be on par with the same PS approach that uses a pre-defined mask. In Chapter 7, we describe in detail how this is accomplished by using image ratios [187] (see (2.88)) and a classical active contour model [41] (see Section 2.5).

## 3.4 Reflectance Parameter Estimation for Large-Scale Scenes

This section discusses related work pertaining to Chapter 8, which aims to address the challenge of bidirectional reflectance distribution function (BRDF) parameter estimation in large-scale scenes. Prior research on material parameter estimation has typically focused on single objects, small scenes, or a single image [32, 47, 56, 74, 107, 111, 112, 127, 128, 134, 142, 198, 204, 243], making it unsuitable for application to larger environments. Our work seeks to predict BRDF parameters for every object in large-scale

scenes using calibrated video input, a corresponding geometry reconstruction, and an object instance segmentation. Other approaches aimed at the same objective have attempted to address this challenge by either restricting material parameter estimation to diffuse empty rooms [253] or by deploying a full path tracing framework [18, 162].

**Zhang *et al.* [253].** The study [253] involves the estimation of materials in room-sized scenes, based on a geometric reconstruction of a non-empty room and the corresponding camera frames. Prior to material estimation, the authors undertake a number of preparatory tasks. First, they perform camera calibration with respect to gamma correction, exposure, and white balance, assuming a purely diffuse scene, resulting in a linear camera response that is then employed for material estimation. In addition, they identify the architectural reconstruction (wall, floor, and ceiling) of the scene, assuming a Manhattan World, whereby all objects are perpendicular to each other along a certain dimension. The authors semi-automatically locate emitters within the scene, enabling the estimation of the intensity of each emitter. The radiometrically calibrated camera frames are then used in conjunction with the architectural reconstruction to estimate the albedo and emitter intensities of the scene, with the assumption that each wall, floor, and ceiling has a constant albedo.
Although the method [253] enables the reconstruction of material parameters in room-sized scans, it is deficient in two important aspects required for realistic re-rendering of scenes. First, the assumption of complete diffuse reflection precludes the accurate modeling of specular reflections, resulting in potentially unrealistic and unfaithful reconstructions. Second, the assumption of constant albedo per wall, ceiling, and floor, restricts the applicability of this approach to scenes that conform to this assumption, precluding spatially varying materials such as a parquet floor.

**Azinovic *et al.* [18].** An approach that removes the constraint of a purely diffuse scene is proposed in [18]. They introduce an inverse path tracer to optimize material parameters and emissivity in a room-sized scene, using a similar input to that of [253] *i.e.*, geometry and calibrated (intrinsically and extrinsically) image captures of the scene. In addition, object instance segmentation is required to optimize for constant emissivity, albedo/baseColor, roughness, and specular parameters per object in the scene. The authors use the Disney BRDF similar to Section 2.3.2.2 to parameterize the reflectance and mathematically formulate light transport in terms of a path integral [226]. Monte Carlo path tracing is nested into a stochastic gradient descent (SGD) optimization scheme using Adam [119] to invert this process and solve for constant material and emissivity parameters, with an $L_1$ regularization term for the emission to increase robustness. After convergence, the authors relax the assumption of constant albedo per object and

perform a second optimization step, where they only optimize for the albedo per face of the mesh. The approach is evaluated mostly on synthetic data that conforms to the assumptions of constant material parameters per scene and a sparse set of non-diffuse material effects. Some evaluation is also conducted on real-world data from the Matterport3D dataset [42], which shows mostly diffuse scenes.
Their method demonstrates promising results in estimating emissive and reflective properties, as evidenced by synthetic experiments. However, when using real-world data, residual noise from the Monte Carlo rendering may contaminate re-rendering or albedo estimates, or the reflectance parameters may lack spatial details, thus hindering the faithful reconstruction of the scene. Moreover, the burden falls on the user to provide a "good" set of input images to estimate emissivity and reflectance parameters. Ideally, the images should capture specular highlights of non-diffuse surfaces; otherwise, these objects may be mistaken as diffuse if no highlights are visible. The authors use only a sparse set of one to three input views for a room-size scene. While it is preferred to minimize the number of images for computational efficiency, selecting only a few views that capture a significant portion of the room, including visible specularity for every non-diffuse object, can be challenging. The authors report a few minutes of computation time if a single image is used, but this can increase to $12$–$24$ hours if more images are employed [2].

**Nimier-David *et al.* [162].** An alternative approach to estimating material parameters and emissivity in large-scale scenes from a given geometry, object instance segmentation and posed images is presented in [162][5]. This approach, which is similar to [18], uses a differentiable path tracer [163] to optimize for emissivity, albedo/baseColor, roughness, and specular parameter of the Disney BRDF (Section 2.3.2.2) using the Adam optimizer [119]. The object segmentation is used in a similar fashion to [18], where a constant value of emissivity, roughness, and specular parameter per object is assumed. To recover high-resolution textures, a UV mapping is used instead of directly optimizing emissivity and BRDF parameters at the mesh's vertices, as done in [18]. The optimization variables are estimated directly in texture space, where a set of uniformly distributed texels[6] is sampled, projected onto the mesh using the inverse UV mapping, and a set of camera views is used for which the camera's frustum captures the corresponding projected texels. This approach automates the manual choosing of "good" input images, although the number of views used is not specified. As the problem is highly non-convex and ill-posed, additional steps are taken to increase numerical stability, including initializing the baseColor/albedo with the median texture from our work presented in this thesis in

[5] It should be noted that [162] appeared around the same time as the work presented in this thesis in Chapter 8, which is why it is not specifically mentioned there.

[6] A texel is a pixel in a texture map.

Chapter 8, using gradient clamping, and preventing optimization updates at zero gradients (which would be updated due to Adam's momentum [119]). A coarse-to-fine update scheme in texture space is also implemented for robustness, where low-resolution textures are optimized first and resolution is gradually increased to $4K$ over the course of optimization.
The research conducted by [162] addresses some of the challenges encountered in [18], such as non-constant albedo estimation and the need for a more automated image selection process. However, a limitation of the proposed approach, similar to that of [18], is the absence of a guarantee that the input images always exhibit specular highlights of non-diffuse objects, despite the use of more than one to three images. As a consequence, the algorithm may yield diffuse BRDF estimates for objects with non-diffuse behavior. Additionally, the high computational complexity of the differentiable path tracer leads to an execution time of $12$ hours.

**Summary.** To summarize, prior work [18, 253] has made certain assumptions, such as a purely diffuse scene, manually selected input images, or a constant set of material parameters per object. While some of these issues have been addressed in a simultaneously published work [162], challenges remain regarding runtime and selecting appropriate input images. Our approach assumes knowledge of emissive objects, a reasonable assumption given the availability of a 3D scan and an object instance segmentation[7]. We estimate material parameters for each object in the scene, including albedo values for each surface point and a set of constant, non-diffuse parameters. In addition, our algorithm introduces a novel method for selecting input images per object that reduces computational complexity and ensures visible specularities. The resulting algorithm processes full, complex scenes with multiple input images in a matter of minutes. Further details on this approach are discussed in Chapter 8.

This chapter provided an overview of the relevant related work that serves as the foundation for the core publications used in this thesis. For each chapter in Part II, we discussed the state-of-the-art, the associated algorithms, their respective advantages, and limitations. While we have alluded to the contributions in these sections, we provide a more in-depth analysis of each paper's individual contribution in the following chapter.

[7]Usually, an object instance segmentation comes along with the classes that are present in the scene *e.g.*, wall, book, lamp. In such a case, one can verify emissive objects by the class they belong to.

# Chapter 4

# Contributions

This chapter provides an overview of the contributions that led to high-quality reconstructions of single objects and room-scale scenes achieved through solving *physically-based inverse problems*. We present a comprehensive list of peer-reviewed publications, of which four form the basis of this cumulative thesis. Subsequently, we provide a detailed description of the contributions of these works in the following section.

## 4.1 List of Publications

The contributions of this thesis stem from four peer-reviewed papers [2, 4, 5, 6], which are the result of fruitful collaborations with esteemed researchers such as Daniel Cremers, Yvain Quéau, Tao Wu, Thomas Möllenhoff, Zhenzhang Ye, Maolin Gao, Thomas Whelan, Simon Green, Michael Goesele, Daniel Andersen, Alan Oursland, and Richard Newcombe. For a list of peer-reviewed articles that have contributed to this dissertation, refer to Table 4.1. Additionally, Table 4.1 also includes other co-authored publications, which are not included in the contributions of this thesis. Notably, all publications were accepted in highly regarded peer-reviewed international conferences or journals.

The research detailed in [4] was performed during a stay as visiting Ph.D. student at the GREYC laboratory, Caen, France. The study [2] was completed as part of a research project during a research internship at Meta[1] Reality Labs Research, Cork, Ireland.
The findings presented in [3] are an extension of the previously published works [5, 7] and provide a more comprehensive and detailed analysis. Furthermore, the articles [1, 3, 8] include outcomes of master's theses that were supervised or co-supervised by the author. Several works were presented as spotlights [4, 5, 8, 9] or full oral presentations [7, 11]. Additionally, the conference paper [11] was invited to a special issue, and its extended version was published in a journal [10].

[1] At the time of the internship, Meta was known as Facebook.

S. Peng, B. Haefner, Y. Quéau, and D. Cremers. Depth Super-Resolution Meets Uncalibrated Photometric Stereo. In *International Conference on Computer Vision (ICCV) Workshops*, 2017 [7]

B. Haefner, Y. Quéau, T. Möllenhoff, and D. Cremers. Fight ill-posedness with ill-posedness: Single-shot variational depth super-resolution from shading. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018 [5] (Chapter 5)

B. Haefner, Z. Ye, M. Gao, T. Wu, Y. Quéau, and D. Cremers. Variational Uncalibrated Photometric Stereo under General Lighting. In *International Conference on Computer Vision (ICCV)*, 2019 [6] (Chapter 6)

B. Haefner, Y. Quéau, and D. Cremers. Photometric Segmentation: Simultaneous Photometric Stereo and Masking. In *International Conference on 3D Vision (3DV)*, 2019 [4] (Chapter 7)

M. Brahimi, Y. Quéau, B. Haefner, and D. Cremers. *On the Well-Posedness of Uncalibrated Photometric Stereo Under General Lighting*. In *Advances in Photometric 3D-Reconstruction*. J.-D. Durou, M. Falcone, Y. Quéau, and S. Tozza, editors. Springer International Publishing, 2020, pages 147–176 [1]

B. Haefner, S. Peng, A. Verma, Y. Quéau, and D. Cremers. Photometric Depth Super-Resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 42(10):2453–2464, 2020 [3]

L. Sang, B. Haefner, and D. Cremers. Inferring Super-Resolution Depth from a Moving Light-Source Enhanced RGB-D Sensor: A Variational Approach. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020 [8]

B. Haefner, S. Green, A. Oursland, D. Andersen, M. Goesele, D. Cremers, R. Newcombe, and T. Whelan. Recovering Real-world Reflectance Properties and Shading from HDR Imagery. In *International Conference on 3D Vision (3DV)*, 2021 [2] (Chapter 8)

Z. Ye, B. Haefner, Y. Quéau, T. Möllenhoff, and D. Cremers. Sublabel-Accurate Multilabeling Meets Product Label Spaces. In *German Conference on Pattern Recognition (GCPR)*, 2021 [11]

Z. Ye, B. Haefner, Y. Quéau, T. Möllenhoff, and D. Cremers. A Cutting-Plane Method for Sublabel-Accurate Relaxation of Problems with Product Label Spaces. *International Journal of Computer Vision (IJCV)*, 2022 [10]

L. Sang, B. Haefner, X. Zuo, and D. Cremers. High-Quality RGB-D Reconstruction via Multi-View Uncalibrated Photometric Stereo and Gradient-SDF. in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2023 [9]

Table 4.1: Full list in chronological order of peer-reviewed publications done within the course of this thesis. The four publications that specifically contribute to this cumulative dissertation are highlighted in black, with additional references to the respective chapters within this work. Other published works that are not part of this thesis are marked in gray for clarity.

## 4.2 Major Contributions

This section provides a comprehensive review of the major contributions included in this thesis. First, we present a contribution that employs Shape-from-Shading (SfS) to recover high-quality super-resolved depth maps from a single-shot RGB-D image pair of objects with piecewise constant reflectance, as detailed in [5]. Next, we describe our work on photometric stereo (PS) under unknown illumination, which achieves state-of-the-art results under a general lighting scenario, as presented in [6]. This is followed by a novel approach to automatically mask an object for photometric stereo (PS)-based problems, described in [4], which circumvents the tedious procedure of manual preprocessing. Finally, we present our work on estimating complex material parameters and shading of large-scale scenes from high dynamic range (HDR) images, which is detailed in [2].

### 4.2.1 Single-Shot Depth Super-Resolution from Shading

Chapter 5 introduces a variational formulation for upsampling a low-resolution (LR) depth map to the same size as its corresponding color image in the context of RGB-D sensors. The approach tackles two difficult problems simultaneously, namely depth super-resolution (SR) and Shape-from-Shading (SfS), both of which are inherently ill-posed. The proposed method utilizes the high-frequency information in the high-resolution RGB image to disambiguate depth SR and the low-frequency information in the LR depth image to disambiguate SfS. Through the joint numerical solution of these problems, the method achieves state-of-the-art results for both tasks. Compared to other approaches in the literature, such as those presented in [164, 242, 246], the proposed method yields crisp, super-resolved depth maps with higher geometric detail.

### 4.2.2 Uncalibrated Photometric Stereo under General Lighting

In Chapter 6, we present an end-to-end, transparent variational problem to solve uncalibrated photometric stereo (UPS) under general lighting conditions. Unlike other methods that optimize for a set of possibly non-integrable normals, we directly optimize for a depth map, ensuring integrability and making our uncalibrated photometric stereo (UPS) setup well-posed [1]. We use Cauchy's M-estimator and Huber-total variation (TV) regularization to make our approach robust against outliers like non-diffuse effects, inter-reflections, or other inaccuracies in the used model. To solve the presented variational problem, we propose a lagged block coordinate descent algorithm. We introduce a novel minimal-surface-based initialization, which effectively helps to recover

state-of-the-art results. Our method outperforms [7, 68, 154] and achieves 2–3× better geometric accuracy in the reconstructions.

### 4.2.3 Simultaneous Photometric Stereo and Masking

In Chapter 7, the aim is to simplify PS approaches by jointly solving 3D reconstruction and 2D masking using a variational formulation. PS data requires both multiple images *and* the object's mask, which is usually generated through a time-consuming manual segmentation procedure [220] as a preprocessing step. The proposed approach eliminates the need for this step by simultaneously solving PS and masking. Unlike other methods that automate the segmentation step resulting in under- or over-segmentation [161], our method generates the best segmentation results by taking the underlying PS problem into account. Furthermore, our method produces a reconstructed surface of higher quality compared to a reconstructed surface without prior segmentation.

### 4.2.4 Recovering Reflectance and Shading From HDR Imagery

The method proposed in Chapter 8 presents a novel approach for recovering material parameters and shading from a set of calibrated HDR video frames and a geometric reconstruction of large-scale scenes. First, we reconstruct diffuse HDR textures of the scene using a novel running median approximation. Simultaneously, we additively split the material in its diffuse and non-diffuse parts, enabling a separate optimization of both. The diffuse material is estimated by Monte Carlo ray tracing, sampling the incident illumination at each point in the scene, thus allowing us to factor the diffuse HDR texture into albedo and shading. Next, we introduce a novel algorithm to automate the selection of target frames. This results in a reduction of computational costs in the subsequent step, as well as an increased likelihood of capturing specular observations. The selected target frames are then used to estimate the non-diffuse material via a ray-tracing-based grid search method with nested least-squares optimization. Our approach effectively leverages HDR data and surpasses similar works [18] in producing high-quality, photo-realistic reconstructions of large-scale scenes.

# Part II

# Physically-Based Inverse Problems

# Chapter 5

# Fight Ill-Posedness with Ill-Posedness: Single-Shot Variational Depth Super-Resolution from Shading

## Copyright


Bjoern Haefner, Yvain Quéau, Thomas Möllenhoff, and Daniel Cremers
**Fight Ill-Posedness with Ill-Posedness: Single-Shot Variational Depth Super-Resolution from Shading**
*2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*
DOI: 10.1109/CVPR.2018.00025

## Individual contributions

Leading role in realizing the scientific project.

| | |
|---|---|
| Problem definition | *significantly contributed* |
| Literature survey | *helped* |
| Implementation | *significantly contributed* |
| Experimental evaluation | *significantly contributed* |
| Preparation of the manuscript | *contributed* |

In accordance with the *IEEE Thesis / Dissertation Reuse Permissions*, we include the accepted version of the original publication [5] in the following.

# Fight ill-posedness with ill-posedness: Single-shot variational depth super-resolution from shading

Bjoern Haefner Yvain Quéau Thomas Möllenhoff Daniel Cremers
Department of Informatics, Technical University of Munich, Germany
{bjoern.haefner,yvain.queau,thomas.moellenhoff,cremers}@tum.de

## Abstract

*We put forward a principled variational approach for up-sampling a single depth map to the resolution of the companion color image provided by an RGB-D sensor. We combine heterogeneous depth and color data in order to jointly solve the ill-posed depth super-resolution and shape-from-shading problems. The low-frequency geometric information necessary to disambiguate shape-from-shading is extracted from the low-resolution depth measurements and, symmetrically, the high-resolution photometric clues in the RGB image provide the high-frequency information required to disambiguate depth super-resolution.*

## 1. Introduction

RGB-D sensors have become very popular for 3D-reconstruction, in view of their low cost and ease of use. They deliver a colored point cloud in a single shot, but the resulting shape often misses thin geometric structures. This is due to noise, quantisation and, more importantly, the coarse resolution of the depth map. However, super-resolution of a solitary depth map without additional constraint is an ill-posed problem.

In comparison, the quality and resolution of the companion RGB image are substantially better. For instance, the Asus Xtion Pro Live device delivers $1280 \times 1024$ px$^2$ RGB images, but only up to $640 \times 480$ px$^2$ depth maps. Therefore, it seems natural to rely on color to refine depth. Yet, retrieving geometry from a single color image is another ill-posed problem, called shape-from-shading. Besides, combining it with depth clues requires the RGB and depth images to have the same resolution.

The resolution of the depth map thus remains a limiting factor in single-shot RGB-D sensing. This work aims at breaking this barrier by jointly refining and upsampling the depth map using shape-from-shading. In other words, **we fight the ill-posedness of single depth image super-resolution using shape-from shading, and vice-versa**.

Input: one low-resolution depth map and its companion high-resolution image — Our result

Figure 1: We carry out single-shot depth super-resolution for commodity RGB-D sensors, using shape-from-shading. By combining low-resolution depth (left) and high-resolution color clues (middle), detail-preserving super-resolution is achieved (right). All figures best viewed in the electronic version.

The rest of this paper is organized as follows. Section 2 reviews the single depth image super-resolution and shape-from-shading problems, in order to motivate their joint solving in the context of RGB-D sensing. Section 3 then introduces a principled Bayesian approach to joint depth super-resolution and shape-from shading. This yields a nonconvex variational problem which is solved using a dedicated ADMM algorithm. Our approach is evaluated against a broad variety of real-world datasets in Section 4, and our conclusions are eventually drawn in Section 5.

## 2. Motivation and related work

Let us first recall the ambiguities arising in single depth image super-resolution and in shape-from-shading, and how they have been handled in the literature.

### 2.1. Ill-posedness in single depth image super-resolution

A depth map is a function which associates to each 2D point of the image plane, the third component of its conjugate 3D-point, relatively to the camera coordinate system. Depth sensors provide out-of-the-box samples of the depth map over a discrete low-resolution rectangular 2D grid $\Omega_{\text{LR}} \subset \mathbb{R}^2$. We will denote by $z_0 : \Omega_{\text{LR}} \to \mathbb{R}$, $p \mapsto z_0(p)$ such a mapping between a pixel $p$ and the measured depth value $z_0(p)$.

Due to hardware constraints, the depth observations $z_0$ are limited by the resolution of the sensor (*i.e.*, the number of pixels in $\Omega_{\text{LR}}$). The single depth image super-resolution problem consists in estimating a high-resolution depth map $z : \Omega_{\text{HR}} \to \mathbb{R}$ over a larger domain $\Omega_{\text{HR}} \supset \Omega_{\text{LR}}$, which coincides with the low-resolution observations $z_0$ over $\Omega_{\text{LR}}$ once it is downsampled. Following [14], this can be formally written as

$$z_0 = Kz + \eta_z. \tag{1}$$

In (1), $K : \mathbb{R}^{\Omega_{\text{HR}}} \to \mathbb{R}^{\Omega_{\text{LR}}}$ is a linear operator combining warping, blurring and downsampling [55]. It can be calibrated beforehand, hence assumed to be known, see for instance [44]. As for $\eta_z$, it stands for the realisation of some stochastic process representing measurement errors, quantisation, etc.

Single depth image super-resolution requires solving Equation (1) in terms of the high-resolution depth map $z$. However, $K$ in (1) maps from a high-dimensional space $\Omega_{\text{HR}}$ to a low-dimensional one $\Omega_{\text{LR}}$, hence it cannot be inverted. Single depth image (blind) super-resolution is thus an ill-posed problem, as there exist infinitely many choices for interpolating between observations, as sketched in Figure 2. Therefore, one must find a way to constrain the problem, as well as to handle noise. This can be achieved by adding observations obtained from different viewing angles [20, 40, 53], but in this work we rather target single-shot applications.

When the input consists in a solitary depth map, disambiguation can be carried out by introducing a smoothness prior on the high-resolution depth map, a strategy which has led to a number of variational approaches, see for instance [55]. More recently, several machine learning approaches have been put forward, which essentially rely on a dictionary of low- and high-resolution depth or edge patches [38, 58]. To avoid resorting to a database, such a dictionary can be constructed from a single depth image by looking for self-similarities [27, 34]. Nevertheless, learning-based depth super-resolution methods remain prone to over-fitting, an issue which has been specifically tackled in [59]. Over-fitting can also be avoided by combining the respective benefits of machine learning and variational approaches [17, 50].

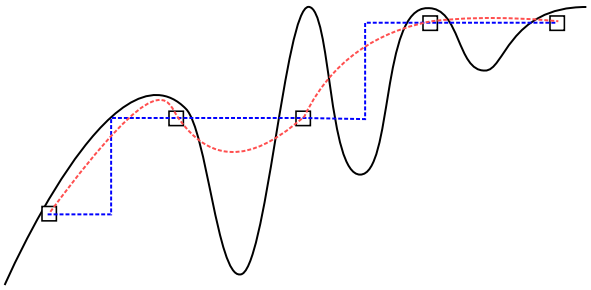


Figure 2: There exist infinitely many ways (dashed lines) to interpolate between low-resolution depth samples (rectangles). Our disambiguation strategy builds upon shape-from-shading applied to the companion high-resolution color image (*cf.* Figure 3), in order to resurrect the fine-scale geometric details of the genuine surface (solid line).

In the RGB-D framework, a high-resolution color image is also available. It can be used as a "guide" to interpolate missing depth values. Several methods were thus proposed to coalign the depth edges in the super-resolved map with edges of the given high-resolution color image [11, 16, 44, 60]. Yet, such approaches only consider sparse features in the high-resolution data, although the whole color image actually conveys shape clues. Indeed, brightness is directly related to the local orientation, hence a photometric approach to depth super-resolution for RGB-D sensors should be feasible and permit to recover fine-scale geometric details. There is, however, surprisingly few works in that direction: to the best of our knowledge, this has been achieved only in [37, 45], but these methods rely on a sequence of images acquired under varying lighting, hence they do not tackle the single-shot problem.

### 2.2. Ill-posedness in shape-from-shading

Shape-from-shading [25] is another classical inverse problem which aims at inferring shape from a single graylevel or color image of a scene. It consists in inverting an image formation model relating the image irradiance $I$ to the scene radiance $\mathcal{R}$, which depends on the surface shape (represented here by the depth map $z$), the incident lighting $l$ and the surface reflectance $\rho$:

$$I = \mathcal{R}(z|l, \rho) + \eta_I, \tag{2}$$

with $\eta_I$ the realisation of a stochastic process standing for noise, quantisation and outliers.

Assuming frontal lighting, uniform Lambertian reflectance, Lipschitz-continuous depth and orthographic projection, solving (2) in terms of the depth map $z$ comes down to solving the eikonal equation [7]

$$|\nabla z| = \sqrt{\frac{1}{I^2} - 1}. \tag{3}$$

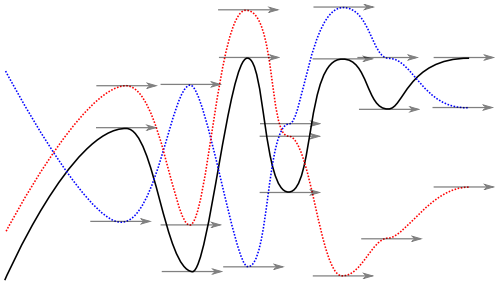


Figure 3: Shape-from-shading suffers from the concave / convex ambiguity: the genuine surface (solid line) and both the surfaces depicted by dashed lines produce the same image, if lit and viewed from above. We put forward low-resolution depth clues (*cf*. Figure 2) for disambiguation.

It is noteworthy that (3) only provides the magnitude of the depth gradient, and not its direction. The local shape is thus unambiguous in singular points (the tangent vectors in Figure 3), but two singular points may either be connected by "going up" or by "going down". This is the well-celebrated concave / convex ambiguity. One out of the infinitely many solutions of (3) can be numerically computed by variational methods [26, 29] or by resorting to the viscosity solution theory [10, 15, 35, 51]. See [6, 12, 62] for further details about numerical shape-from-shading.

Even under the unrealistic assumptions yielding the eikonal shape-from-shading model (3), shape inference is ill-posed. Hence, one may expect that more realistic lighting and reflectance assumptions will add more ambiguities. Several steps in the direction of handling natural lighting have been achieved [28, 31, 49], but they still require the reflectance to be uniform. However, in general both the lighting and the reflectance may be arbitrarily complex. This is nicely visualized in the "workshop metaphor" of Adelson and Pentland [1]: any image can be explained by a flat shape illuminated uniformly but painted in a complex manner, by a white and frontally-lit surface with a complex geometry, or by a white planar surface illuminated in a complex manner. To solve this series of ambiguities, additional constraints must be introduced. Barron *et al*. proposed for this purpose appropriate priors for reflectance (sparsity of the gradients), lighting (spherical harmonics model [4, 48]) and shape (smoothness), and combined them in order to achieve shape, reflectance and illumination from shading [3].

Recently, the shape-from-shading problem has gained new life with the emergence of RGB-D sensors. Indeed, the rough depth map can be used as prior to "guide" shape-from-shading and thus circumvent its ambiguities. This has been achieved in the multi-view setup [39, 63], but also in the single-shot case [9, 22, 42, 43, 57, 61] we tackle in this paper. Still, these methods require the resolutions of the input image and of the depth map to be the same.

### 2.3. Intuitive justification of our proposal

In view of this brief discussion on single depth image super-resolution and shape-from-shading, we conclude that, in the context of RGB-D sensing, the high-frequency information necessary to achieve detail-preserving depth super-resolution could be provided by the photometric data. Similarly, the low-frequency information necessary to disambiguate shape-from-shading could be conveyed by the geometric data. Compare Figures 2 and 3, and see Figure 4. It should thus be possible to achieve joint depth map refinement and super-resolution in a single shot, without resorting to additional data (new viewing angles or illumination conditions, learnt dictionary, etc.). In the next section, we formulate this task as a principled variational problem, by resorting to Bayesian inference.

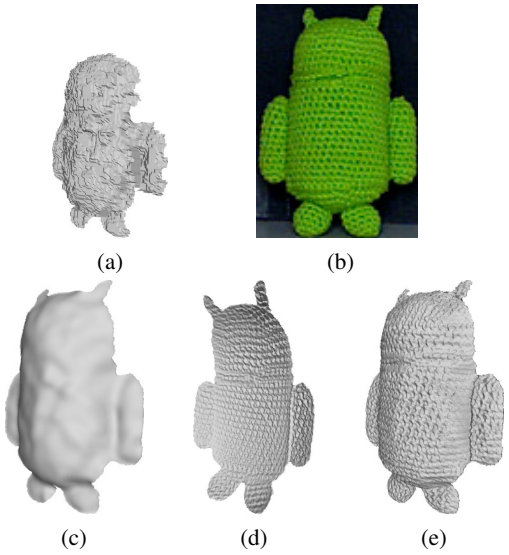



Figure 4: (a-b) Input low-resolution depth and high-resolution color images. (c) Blind super-resolution (achieved by disabling the shape-from-shading term in (18)) cannot hallucinate high-frequency geometric details from (a). (d) Shape-from-shading (achieved by setting $\mu = 0$ in (18)) applied to (b) appropriately recover such thin structures, but it is prone to low-frequency errors. (e) The combination of both techniques yields appropriate restoration of both high- and low-frequency components.

## 3. A variational approach to joint depth super-resolution and shape-from-shading

We formulate shading-based depth super-resolution as the joint solving of (1) (super-resolution) and (2) (shape-from-shading) in terms of the high-resolution depth map $z : \Omega_{\text{HR}} \to \mathbb{R}$, given a low-resolution depth map $z_0 : \Omega_{\text{LR}} \to \mathbb{R}$ and a high-resolution RGB image $I : \Omega_{\text{HR}} \to \mathbb{R}^3$.

We aim at recovering not only a high-resolution depth map which is consistent both with the low-resolution depth measurements and with the high-resolution color data, but also the hidden parameters of the image formation model (2) *i.e.*, the reflectance $\rho$ and the lighting $l$. This can be achieved by maximizing the posterior distribution of the input data which, according to Bayes rule, is given by

$$\mathcal{P}(z,\rho,l|z_0,I) = \frac{\mathcal{P}(z_0,I|z,\rho,l)\,\mathcal{P}(z,\rho,l)}{\mathcal{P}(z_0,I)}, \tag{4}$$

where the numerator is the product of the likelihood with the prior, and the denominator is the evidence, which can be discarded since it plays no role in maximum a posteriori (MAP) estimation. In order to make the independency assumptions as transparent as possible and to motivate the final energy we aim at minimizing (see (18)), we follow in the next subsections David Mumford's approach [41] to derive a variational model from the posterior distribution (4).

### 3.1. Likelihood

Let us start with the first term in the numerator of (4) *i.e.*, the likelihood. By construction of RGB-D sensors, depth and color observations are independent, hence $\mathcal{P}(z_0,I|z,\rho,l) = \mathcal{P}(z_0|z,\rho,l)\,\mathcal{P}(I|z,\rho,l)$. We further assume that the depth observations are independent from the surface reflectance and from the lighting, hence $\mathcal{P}(z_0|z,\rho,l) = \mathcal{P}(z_0|z)$ and thus:

$$\mathcal{P}(z_0,I|z,\rho,l) = \mathcal{P}(z_0|z)\,\mathcal{P}(I|z,\rho,l). \tag{5}$$

Assuming homoskedastic, zero-mean Gaussian noise $\eta_z$ with variance $\sigma_z^2$ in (1), the first factor in (5) writes

$$\mathcal{P}(z_0|z) \propto \exp\left\{-\frac{\|Kz - z_0\|^2_{\ell^2(\Omega_{\text{LR}})}}{2\sigma_z^2}\right\}. \tag{6}$$

Next, we discuss the second factor in (5), by making Equation (2) explicit. In general, the irradiance in channel $\star \in \{R,G,B\}$ writes

$$I_\star = \int_\lambda \int_\omega c_\star(\lambda)\rho(\lambda)\phi(\lambda,\omega)\max\{0,\mathbf{s}(\omega)\cdot\mathbf{n}_z\}\,\mathrm{d}\omega\,\mathrm{d}\lambda + \eta_I, \tag{7}$$

where integration is carried out over all wavelengths $\lambda$ ($\rho$ is the spectral reflectance of the surface and $c_\star$ is the transmission spectrum of the camera in channel $\star$) and all incident lighting directions $\omega$ ($\mathbf{s}(\omega)$ is the unit-length vector pointing towards the light source located in direction $\omega$, and $\phi(\cdot,\omega)$ is the spectrum of this source), and $\mathbf{n}_z$ is the unit-length surface normal (which depends on the underlying depth map $z$). Assuming achromatic lighting i.e., $\phi(\cdot,\omega) := \phi(\omega)$, and using a first-order[1] spherical harmonics approximation of the inner integral, we obtain

$$I = \underbrace{\begin{bmatrix}\int_\lambda c_R(\lambda)\rho(\lambda)\mathrm{d}\lambda\\ \int_\lambda c_G(\lambda)\rho(\lambda)\mathrm{d}\lambda\\ \int_\lambda c_B(\lambda)\rho(\lambda)\mathrm{d}\lambda\end{bmatrix}}_{:=\rho} l\cdot\begin{bmatrix}\mathbf{n}_z\\ 1\end{bmatrix} + \eta_I, \tag{8}$$

with $l \in \mathbb{R}^4$ the achromatic "light vector", $\rho : \Omega_{\text{HR}} \to \mathbb{R}^3$ the albedo (Lambertian reflectance) map, relatively to the camera transmission spectra $\{c_\star\}_{\star\in\{R,G,B\}}$, and $\mathbf{n}_z : \Omega_{\text{HR}} \to \mathbb{S}^2 \subset \mathbb{R}^3$ the field of unit-length surface normals. Assuming perspective projection with focal length $f > 0$ and $\mathbf{p} : \Omega_{\text{HR}} \to \mathbb{R}^2$ the field of pixel coordinates with respect to the principal point, the normal field is given by

$$\mathbf{n}_z = \frac{1}{\sqrt{|f\,\nabla z|^2 + (-z - \mathbf{p}\cdot\nabla z)^2}}\begin{bmatrix} f\,\nabla z \\ -z - \mathbf{p}\cdot\nabla z\end{bmatrix} \tag{9}$$

(see, for instance, [46]).

Assuming that the image noise is homoskedastically Gaussian-distributed with zero-mean and covariance matrix $\mathrm{Diag}(\sigma_I^2,\sigma_I^2,\sigma_I^2)$, we obtain

$$\mathcal{P}(I|z,\rho,l) \propto \exp\left\{-\frac{\|(l\cdot\mathbf{m}_{z,\nabla z})\,\rho - I\|^2_{\ell^2(\Omega_{\text{HR}})}}{2\sigma_I^2}\right\}, \tag{10}$$

where, according to (8) and (9), $\mathbf{m}_{z,\nabla z}$ is a $\Omega_{\text{HR}} \to \mathbb{R}^4$ vector field defined as

$$\mathbf{m}_{z,\nabla z} = \begin{bmatrix} \dfrac{f\,\nabla z}{\sqrt{|f\,\nabla z|^2 + (-z-\mathbf{p}\cdot\nabla z)^2}} \\ \dfrac{-z-\mathbf{p}\cdot\nabla z}{\sqrt{|f\,\nabla z|^2 + (-z-\mathbf{p}\cdot\nabla z)^2}} \\ 1\end{bmatrix}. \tag{11}$$

### 3.2. Priors

We now consider the second factor in the numerator of (4) *i.e.*, the prior distribution. We assume that depth, reflectance and lighting are independent (independence of reflectance from depth and lighting follows from the Lambertian assumption, and independence of lighting from depth follows from the distant-light assumption required to derive the spherical harmonics model (8), see [4, 48]). This implies

$$\mathcal{P}(z,\rho,l) = \mathcal{P}(z)\,\mathcal{P}(\rho)\,\mathcal{P}(l). \tag{12}$$

Since lighting has already been modeled as a low-frequency phenomenon for the sake of expliciting the image formation model (8), we do not need to introduce any other prior $\mathcal{P}(l)$ and thus we use an improper prior

$$\mathcal{P}(l) = \text{constant}. \tag{13}$$

[1] The whole proposed method is straightforward to extend to second-order spherical harmonics. However we did not observe substantial improvement with this extension, hence we discuss only the first-order case, which can capture more than 85% of natural illumination [18].

Regarding the depth map $z$, we follow the recent work [21] and opt for a minimal surface prior. Remark that

$$\mathrm{d}\mathcal{A}_{z,\nabla z} = \frac{z}{f^2}\sqrt{|f\,\nabla z|^2 + \left(-z - \mathbf{p}\cdot\nabla z\right)^2} \tag{14}$$

is a $\Omega_{\mathrm{HR}} \to \mathbb{R}$ scalar field which maps each pixel to the area of the corresponding surface element. Thus $\|\mathrm{d}\mathcal{A}_{z,\nabla z}\|_{\ell^1(\Omega_{\mathrm{HR}})}$ is the total surface area and the minimal surface prior writes

$$\mathcal{P}(z) \propto \exp\left\{-\frac{\|\mathrm{d}\mathcal{A}_{z,\nabla z}\|_{\ell^1(\Omega_{\mathrm{HR}})}}{\alpha}\right\}, \tag{15}$$

with $\alpha > 0$ a free parameter controlling smoothness.

According to the Retinex theory [33], the reflectance $\rho$ can be assumed piecewise constant. This yields a Potts prior

$$\mathcal{P}(\rho) \propto \exp\left\{-\frac{\|\nabla\rho\|_{\ell^0(\Omega_{\mathrm{HR}})}}{\beta}\right\}, \tag{16}$$

with $\beta > 0$ a scale parameter, and $\|\cdot\|_{\ell^0}$ an abusive notation for the length of the discontinuity set:

$$\|\nabla\rho\|_{\ell^0(\Omega_{\mathrm{HR}})} = \sum_{p\in\Omega_{\mathrm{HR}}} \begin{cases} 0, & \text{if } |\nabla\rho(p)|_2 = 0, \\ 1, & \text{otherwise,} \end{cases} \tag{17}$$

where $|\cdot|_2$ is the Euclidean norm in $\mathbb{R}^6$.

### 3.3. Variational formulation

Replacing the maximisation of the posterior distribution (4) by the minimisation of its negative logarithm, combining Equations (4)–(6), (10), (12)–(16), and neglecting the additive constants, we end up with the variational model

$$\begin{aligned}\min_{\substack{\rho\colon \Omega_{\mathrm{HR}}\to\mathbb{R}^3 \\ l\in\mathbb{R}^4 \\ z\colon \Omega_{\mathrm{HR}}\to\mathbb{R}}} & \|(l\cdot\mathbf{m}_{z,\nabla z})\,\rho - I\|^2_{\ell^2(\Omega_{\mathrm{HR}})} + \mu\,\|Kz - z_0\|^2_{\ell^2(\Omega_{\mathrm{LR}})} \\ & + \nu\,\|\mathrm{d}\mathcal{A}_{z,\nabla z}\|_{\ell^1(\Omega_{\mathrm{HR}})} + \lambda\,\|\nabla\rho\|_{\ell^0(\Omega_{\mathrm{HR}})}\,,\end{aligned} \tag{18}$$

with the following definitions of the weights:

$$\mu = \frac{\sigma_I^2}{\sigma_z^2}, \quad \nu = \frac{2\sigma_I^2}{\alpha} \text{ and } \lambda = \frac{2\sigma_I^2}{\beta}. \tag{19}$$

### 3.4. Numerical solution

We now describe an algorithm for effectively solving the variational problem (18), which is both nonsmooth and nonconvex. In order to tackle the nonlinear dependency upon the depth and its gradient arising from shape-from-shading and minimal surface regularisation, we follow [47] and introduce an auxiliary variable $\theta := (z, \nabla z)$, then rewrite (18) as a constrained optimisation problem:

$$\begin{aligned}\min_{\substack{\rho\colon \Omega_{\mathrm{HR}}\to\mathbb{R}^3 \\ l\in\mathbb{R}^4 \\ z\colon \Omega_{\mathrm{HR}}\to\mathbb{R} \\ \theta\colon \Omega_{\mathrm{HR}}\to\mathbb{R}^3}} & \|(l\cdot\mathbf{m}_{\theta})\,\rho - I\|^2_{\ell^2(\Omega_{\mathrm{HR}})} + \mu\,\|Kz - z_0\|^2_{\ell^2(\Omega_{\mathrm{LR}})} \\ & + \nu\,\|\mathrm{d}\mathcal{A}_{\theta}\|_{\ell^1(\Omega_{\mathrm{HR}})} + \lambda\,\|\nabla\rho\|_{\ell^0(\Omega_{\mathrm{HR}})} \\ \text{s.t. } & \theta = (z, \nabla z).\end{aligned} \tag{20}$$

We then use a multi-block variant of ADMM [5, 13, 19] to solve (20)[2]. Given the current estimates $(\rho^{(k)}, l^{(k)}, \theta^{(k)}, z^{(k)})$ at iteration $(k)$, the variables are updated according to the following sweep:

$$\begin{aligned}\rho^{(k+1)} = \operatorname*{argmin}_{\rho} & \left\|\left(l^{(k)}\cdot\mathbf{m}_{\theta^{(k)}}\right)\rho - I\right\|^2_{\ell^2(\Omega_{\mathrm{HR}})} \\ & + \lambda\,\|\nabla\rho\|_{\ell^0(\Omega_{\mathrm{HR}})}\,,\end{aligned} \tag{21}$$

$$l^{(k+1)} = \operatorname*{argmin}_{l} \left\|\left(l\cdot\mathbf{m}_{\theta^{(k)}}\right)\rho^{(k+1)} - I\right\|^2_{\ell^2(\Omega_{\mathrm{HR}})}, \tag{22}$$

$$\begin{aligned}\theta^{(k+1)} = \operatorname*{argmin}_{\theta} & \left\|\left(l^{(k+1)}\cdot\mathbf{m}_{\theta}\right)\rho^{(k+1)} - I\right\|^2_{\ell^2(\Omega_{\mathrm{HR}})} \\ & + \nu\|\mathrm{d}\mathcal{A}_{\theta}\|_{\ell^1(\Omega_{\mathrm{HR}})} + \frac{\kappa}{2}\left\|\theta - (z,\nabla z)^{(k)} + u^{(k)}\right\|^2_{\ell^2(\Omega_{\mathrm{HR}})},\end{aligned} \tag{23}$$

$$\begin{aligned}z^{(k+1)} = \operatorname{argmin}_{z}\ & \mu\,\|Kz - z_0\|^2_{\ell^2(\Omega_{\mathrm{LR}})} \\ & + \frac{\kappa}{2}\left\|\theta^{(k+1)} - (z,\nabla z) + u^{(k)}\right\|^2_{\ell^2(\Omega_{\mathrm{HR}})},\end{aligned} \tag{24}$$

$$u^{(k+1)} = u^{(k)} + \theta^{(k+1)} - (z^{(k+1)}, \nabla z^{(k+1)}), \tag{25}$$

where $u$ and $\kappa$ are a Lagrange multiplier and a step size, respectively. In our implementation $\kappa$ is determined automatically using the varying penalty procedure [23].

To solve the albedo sub-problem (21) we resort to primal-dual iterations [54]. The lighting update (22) is solved using pseudo-inverse. The $\theta$-update (23) comes down to a series of independent (there is no coupling between neighboring pixels, thanks to the ADMM strategy) nonlinear optimisation problems, which we solve using the implementation [52] of the L-BFGS method [36], using the Moreau envelope of the $\ell^1$ norm to ensure differentiability. The depth update (24) requires solving a large sparse linear least-squares problem, which we tackle using conjugate gradient on the normal equations.

Although the overall optimisation problem (18) is nonconvex, recent works [24, 30, 56] have demonstrated that under mild assumptions on the cost function and small enough step size $\kappa$, nonconvex ADMM converges to a critical point. In practice, we found the proposed ADMM scheme to be stable and always observed convergence. In our experiments we use as initial guess: $\rho^{(0)} = I$, $l^{(0)} = [0, 0, -1, 0]^\top$, $z^{(0)}$ a smoothed (using bilinear filtering) version of a linear interpolation of the low-resolution input $z_0$, $\theta^{(0)} = (z^0, \nabla z^{(0)})$, $u^{(0)} \equiv 0$ and $\kappa^{(0)} = 10^{-4}$. In all our experiments, 10 to 20 global iterations $(k)$ were sufficient to reach convergence, which is evaluated through the relative residual between two successive depth estimates $z^{(k+1)}$ and $z^{(k)}$. On a recent laptop computer with $i7$ processor, such a process requires around one minute (code is implemented in Matlab except the albedo update, which is implemented in CUDA).

[2] Code and dataset is available at https://github.com/BjoernHaefner/DepthSRfromShading.

## 4. Experimental validation

In this section we evaluate our variational approach to joint depth super-resolution and shape-from-shading against challenging synthetic and real-world datasets.

### 4.1. Synthetic data

We first discuss the choice of the parameters involved in the variational problem (18). Although their optimal values can be deduced from the data statistics (see (19)), it can be difficult to estimate such statistics in practice and thus we rather consider $\mu$, $\nu$ and $\lambda$ as tunable hyper-parameters. The formulae in (19) remain however insightful regarding the way these parameters should be tuned.

To select an appropriate set of parameters, we consider a synthetic dataset (the publicly available "Joyful Yell" 3D-shape) which we render under first-order spherical harmonics lighting ($l = [0, 0, -1, 0.2]^\top$) with three different reflectance maps as depicted in Figure 5. Additive zero-mean Gaussian noise with standard deviation $1\%$ that of the original images is added to the high resolution ($640 \times 480$ px$^2$) images. Ground-truth high resolution and input low-resolution ($320 \times 240$ px$^2$) depth maps are rendered from the 3D-model. Non-uniform zero-mean Gaussian noise with standard deviation $10^{-3}$ times the squared original depth value (consistently with the real-world measurements from [32]) is then added to the low-resolution depth map. Quantitative evaluation is carried out by evaluating the root mean squared error (RMSE) between the estimated depth and albedo maps and the ground-truth ones.

Initially, we chose $\mu = \frac{1}{12}$, $\nu = 2$ and $\lambda = 1$. Then, we evaluated the impact of varying each parameter, keeping the others fixed to these values found empirically. Results are shown in Figure 6. Quite logically, $\mu$ should not be set too high otherwise the resulting depth map is as noisy as the input. Low values always allow a good albedo estimation, but the range $\mu \in [10^{-2}, 1]$ seems to provide the most accurate depth maps. Regarding $\lambda$, larger values should be chosen if the reflectance is uniform, but they induce high errors whenever it is not. On the other hand, low values systematically yield high errors since the reflectance estimate absorbs all the shading information (this is the "painter's explanation" in the "workshop metaphor" [1]). In between, the range $\lambda \in [10^{-1}, 10]$ seems to always give reasonable results. Eventually, high values of $\nu$ should be avoided in order to prevent over-smoothing.

Since we chose to disambiguate shape-from-shading by assuming piecewise-constant reflectance, the minimal surface prior plays no role in disambiguation. This explains why low values of $\nu$ should be preferred. Depth regularisation matters only when color cannot be exploited, for instance due to shadows, black reflectance or saturation. This will be better visualised in the real-world experiments.

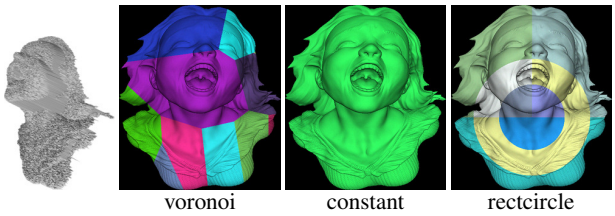



Figure 5: Synthetic dataset used for quantitative evaluation. Left: low-resolution depth map. Right: high-resolution RGB images, rendered using three different albedo maps.

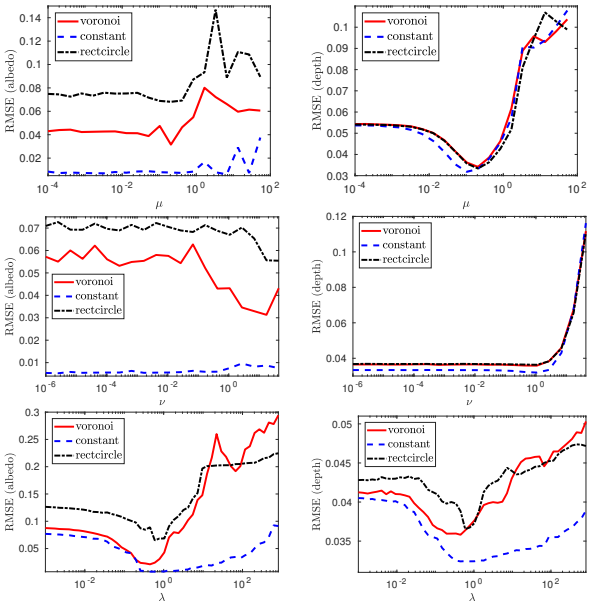



Figure 6: Impact of the parameters $\mu$, $\nu$ and $\lambda$ on the accuracy of the albedo and depth estimates. Based on those experiments, we select the set of parameters $(\mu, \nu, \lambda) = (10^{-1}, 10^{-1}, 2)$ for our experiments.

In Figure 7, we compare our method with two other single-shot ones: a learning-based approach [58] and an image-based one [60]. To emphasise the interest of joint shape-from-shading and super-resolution over shading-based depth refinement using the downsampled image, we also show the results of [43]. For fair comparison with [58], this time we use a scaling factor of 4 for all methods *i.e.*, the depth maps are rendered at $120 \times 160$ px$^2$. To evaluate the recovery of thin structures, we provide the mean angular error with respect to surface normals. The learning-based method can obviously not hallucinate surface details since it does not use the color image. The image-based method does a much better job, but it is largely overcome by shading-based super-resolution.

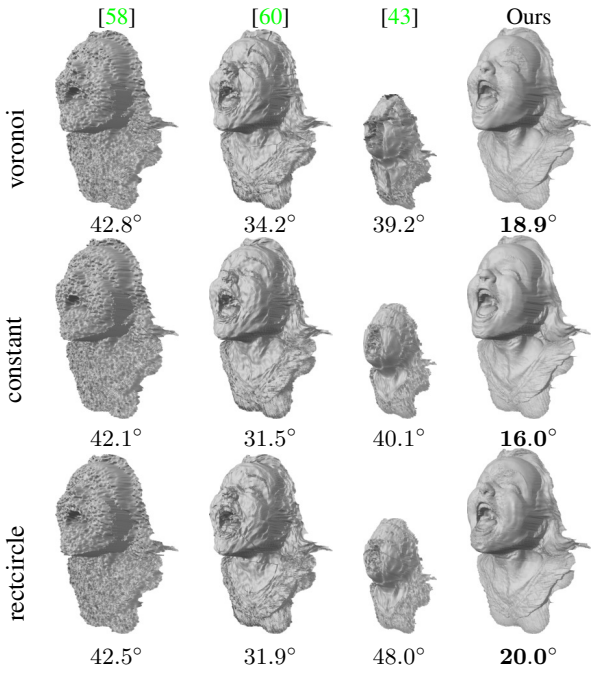



Figure 7: Comparison between learning-based [58], image-based [60] and shading-based (ours) depth super-resolution, as well as shading-based refinement using low-resolution images [43]. Our method systematically outperforms the others (numbers are the mean angular errors on normals).

### 4.2. Real-world data

For real-world experiments, we use the Asus Xtion Pro Live sensor, which delivers $1280 \times 1024$ px$^2$ RGB and $640 \times 480$ px$^2$ depth images at 30 fps. Data are acquired in an indoor office with ambient lighting, and objects are manually segmented from background before processing.

Figures 1, 4, 8, 9, 10 and 13 present real-world results. Combining depth super-resolution and shape-from-shading apparently resolves the low-frequency and high-frequency ambiguities arising in either of the inverse problems. Over-segmentation of reflectance may happen, but this does not seem to impact depth recovery. Whenever color gets saturated or too low, then minimal surface drives super-resolution, which adds robustness. Additional results using depth maps with lower resolution ($320 \times 240$ px$^2$) are presented in Figure 11. Our method only fails when reflectance does not fit the Potts prior, as shown in Figure 12 for an object with smoothly-varying reflectance. It induces bias in the estimated depth such that reflectance based artifacts appear. Handling such cases would require using another prior for the reflectance, or actively controlling lighting. This has already been achieved in RGB-D sensing [2, 8, 45], but it is not compatible with single-shot applications.

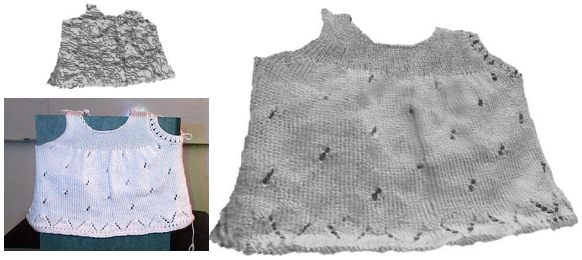


Figure 8: Super-resolution of a "dress". The estimated reflectance map is uniform, hence it is not displayed here.

Figure 9: Super-resolution of a "monkey doll". Fine-scale shape and reflectance structures are nicely recovered.

Figure 10: Super-resolution of "wool balls". Minimal surface drives super-resolution when color gets saturated.

## 5. Conclusion

A variational approach to single-shot depth super-resolution for RGB-D sensors is proposed. It fully exploits the color information in order to guide super-resolution, by resorting to the shape-from-shading technique. Low-resolution depth cues resolve the ambiguities arising in shape-from-shading and, symmetrically, high-resolution photometric clues resolve those of depth super-resolution.

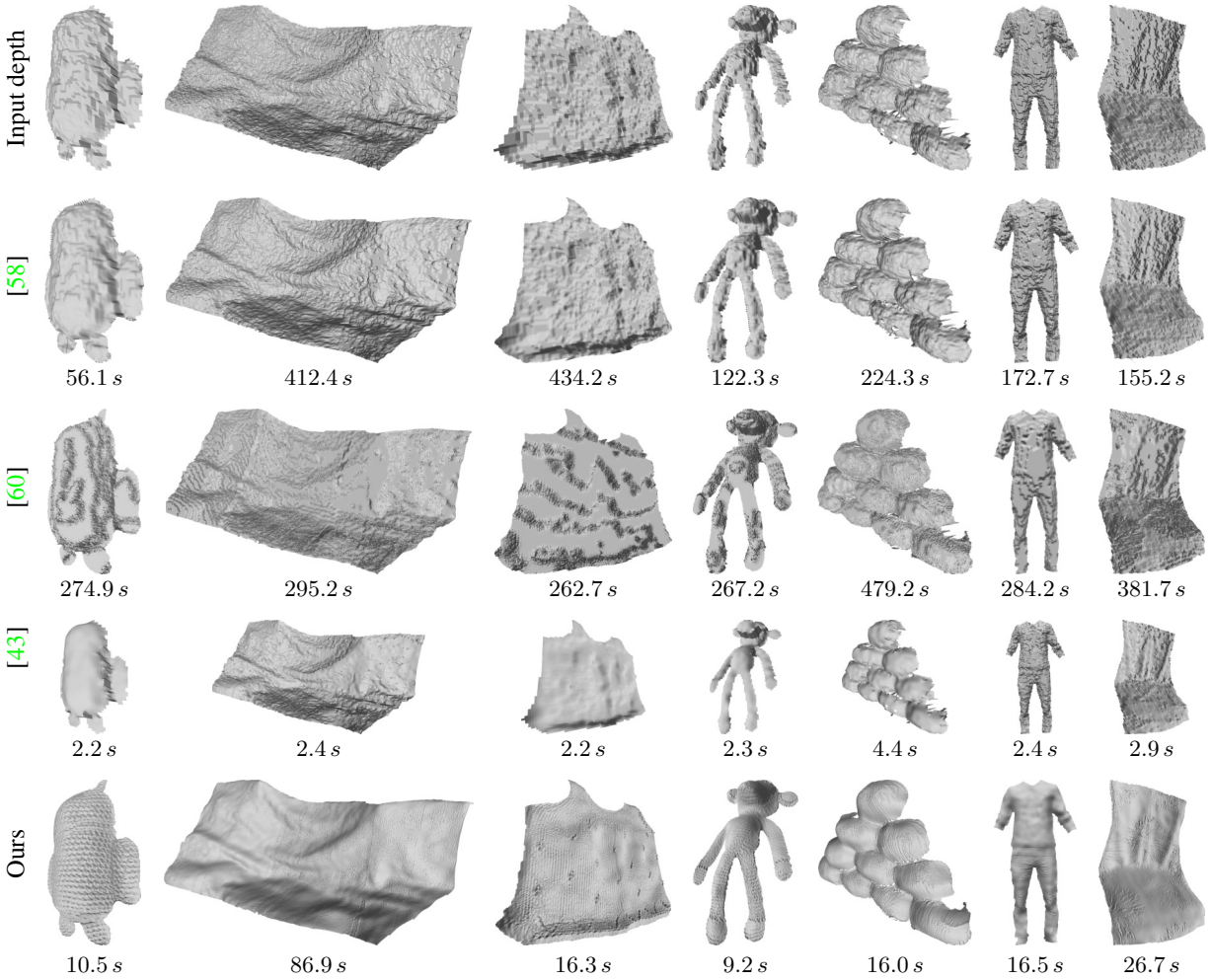



Figure 11: Comparison between our super-resolution method, two others [58, 60] and shading-based depth refinement on the low-resolution images [43]. Our shading-based super-resolution restores the complex geometry the best. Numbers represent runtime in seconds.

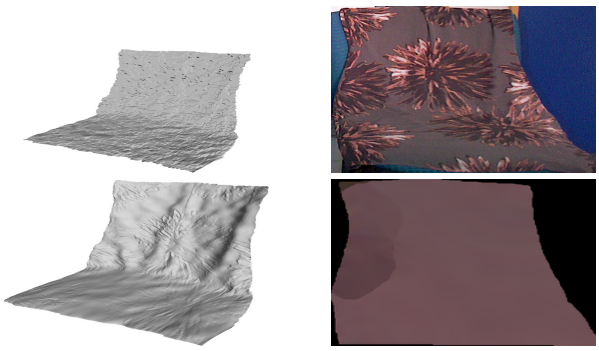


Figure 12: If the pictured object does not match our Potts prior for the reflectance, artifacts appear.

Figure 13: Super-resolution of a "blanket". Despite over-segmentation of the reflectance, thin structures are recovered. Even in black areas without shading information, results remain satisfactory thanks to the minimal surface prior.

## References


[1] E. H. Adelson and A. P. Pentland. *Perception as Bayesian inference*, chapter The perception of shading and reflectance, pages 409–423. Cambridge University Press, 1996. 3, 6

[2] R. Anderson, B. Stenger, and R. Cipolla. Augmenting depth camera output using photometric stereo. In *Proceedings of the IAPR Conference on Machine Vision Applications*, 2011. 7

[3] J. Barron and J. Malik. Shape, illumination, and reflectance from shading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(8):1670–1687, 2015. 3

[4] R. Basri and D. P. Jacobs. Lambertian reflectances and linear subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(2):218–233, 2003. 3, 4

[5] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011. 5

[6] M. Breuß, E. Cristiani, J.-D. Durou, M. Falcone, and O. Vogel. Perspective shape from shading: Ambiguity analysis and numerical approximations. *SIAM Journal on Imaging Sciences*, 5(1):311–342, 2012. 3

[7] A. R. Bruss. The eikonal equation: Some results applicable to computer vision. *Journal of Mathematical Physics*, 23(5):890–896, 1982. 2

[8] A. Chatterjee and V. Madhav Govindu. Photometric refinement of depth maps for multi-albedo objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 933–941, 2015. 7

[9] G. Choe, J. Park, Y.-W. Tai, and I. S. Kweon. Refining geometry from depth sensors using IR shading images. *International Journal of Computer Vision*, 122(1):1–16, 2017. 3

[10] E. Cristiani and M. Falcone. Fast semi-lagrangian schemes for the eikonal equation and applications. *SIAM Journal on Numerical Analysis*, 45(5):1979–2011, 2007. 3

[11] J. Diebel and S. Thrun. An application of Markov random fields to range sensing. In *Advances in Neural Information Processing Systems*, pages 291–298, 2006. 2

[12] J.-D. Durou, M. Falcone, and M. Sagona. Numerical Methods for Shape-from-shading: A New Survey with Benchmarks. *Computer Vision and Image Understanding*, 109(1):22–43, 2008. 3

[13] J. Eckstein and D. P. Bertsekas. On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1):293–318, 1992. 5

[14] M. Elad and A. Feuer. Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images. *IEEE Transactions on Image Processing*, 6(12):1646–1658, 1997. 2

[15] M. Falcone and M. Sagona. An algorithm for the global solution of the shape-from-shading model. In *Proceedings of the International Conference on Image Analysis and Processing*, pages 596–603, 1997. 3

[16] D. Ferstl, C. Reinbacher, R. Ranftl, M. Rüther, and H. Bischof. Image guided depth upsampling using anisotropic total generalized variation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 993–1000, 2013. 2

[17] D. Ferstl, M. Ruther, and H. Bischof. Variational depth superresolution using example-based edge representations. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 513–521, 2015. 2

[18] D. Frolova, D. Simakov, and R. Basri. Accuracy of spherical harmonic approximations for images of Lambertian objects under far and near lighting. In *Proceedings of the European Conference on Computer Vision*, pages 574–587, 2004. 4

[19] R. Glowinski and A. Marroco. Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de Dirichlet non linéaires. *Revue française d'automatique, informatique, recherche opérationnelle. Analyse numérique*, 9(R2):41–76, 1975. 5

[20] B. Goldlücke, M. Aubry, K. Kolev, and D. Cremers. A super-resolution framework for high-accuracy multiview reconstruction. *International Journal of Computer Vision*, 106(2):172–191, 2014. 2

[21] G. Graber, J. Balzer, S. Soatto, and T. Pock. Efficient minimal-surface regularization of perspective depth maps in variational stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 511–520, 2015. 5

[22] Y. Han, J.-Y. Lee, and I. S. Kweon. High Quality Shape from a Single RGB-D Image under Uncalibrated Natural Illumination. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1617–1624, 2013. 3

[23] B. S. He, H. Yang, and S. L. Wang. Alternating direction method with self-adaptive penalty parameters for monotone variational inequalities. *Journal of Optimization Theory and Applications*, 106(2):337–356, 2000. 5

[24] M. Hong, Z.-Q. Luo, and M. Razaviyayn. Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems. *SIAM Journal on Optimization*, 26(1):337–364, 2016. 5

[25] B. K. P. Horn. *Shape From Shading: A Method for Obtaining the Shape of a Smooth Opaque Object From One View*. PhD thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 1970. 2

[26] B. K. P. Horn and M. J. Brooks. The variational approach to shape from shading. *Computer Vision, Graphics, and Image Processing*, 33(2):174–208, 1986. 3

[27] M. Hornácek, C. Rhemann, M. Gelautz, and C. Rother. Depth super resolution by rigid body self-similarity in 3D. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1123–1130, 2013. 2

[28] R. Huang and W. A. P. Smith. Shape-from-shading under complex natural illumination. In *Proceedings of the IEEE International Conference on Image Processing*, pages 13–16, 2011. 3

[29] K. Ikeuchi and B. K. Horn. Numerical shape from shading and occluding boundaries. *Artificial intelligence*, 17(1-3):141–184, 1981. 3

[30] B. Jiang, T. Lin, S. Ma, and S. Zhang. Structured nonconvex and nonsmooth optimization: algorithms and iteration complexity analysis. *arXiv preprint arXiv:1605.02408*, 2016. 5

[31] M. K. Johnson and E. H. Adelson. Shape estimation in natural illumination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2553–2560, 2011. 3

[32] K. Khoshelham and S. O. Elberink. Accuracy and resolution of Kinect depth data for indoor mapping applications. *Sensors*, 12(2):1437–1454, 2012. 6

[33] E. H. Land. The retinex theory of color vision. *Scientific American*, 237(6):108–120, 1977. 5

[34] J. Li, Z. Lu, G. Zeng, R. Gan, and H. Zha. Similarity-aware patchwork assembly for depth image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3374–3381, 2014. 2

[35] P.-L. Lions, E. Rouy, and A. Tourin. Shape-from-shading, viscosity solutions and edges. *Numerische Mathematik*, 64(1):323–353, 1993. 3

[36] D. C. Liu and J. Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical programming*, 45(1):503–528, 1989. 5

[37] Z. Lu, Y.-W. Tai, F. Deng, M. Ben-Ezra, and M. S. Brown. A 3D imaging framework based on high-resolution photometric-stereo and low-resolution depth. *International Journal of Computer Vision*, 102(1-3):18–32, 2013. 2

[38] O. Mac Aodha, N. D. F. Campbell, A. Nair, and G. J. Brostow. Patch based synthesis for single depth image super-resolution. In *Proceedings of the European Conference on Computer Vision*, pages 71–84, 2012. 2

[39] R. Maier, K. Kim, D. Cremers, J. Kautz, and M. Nießner. Intrinsic3d: High-quality 3D reconstruction by joint appearance and geometry optimization with spatially-varying lighting. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 3

[40] R. Maier, J. Stückler, and D. Cremers. Super-resolution keyframe fusion for 3D modeling with high-quality textures. In *Proceedings of the International Conference on 3D Vision*, pages 536–544, 2015. 2

[41] D. Mumford. Bayesian rationale for the variational formulation. In *Geometry-driven diffusion in computer vision*, pages 135–146. 1994. 4

[42] R. Or-El, R. Hershkovitz, A. Wetzler, G. Rosman, A. M. Bruckstein, and R. Kimmel. Real-time depth refinement for specular objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4378–4386, 2016. 3

[43] R. Or-El, G. Rosman, A. Wetzler, R. Kimmel, and A. Bruckstein. RGBD-Fusion: Real-Time High Precision Depth Recovery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5407–5416, 2015. 3, 6, 7, 8

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

[45] S. Peng, B. Haefner, Y. Quéau, and D. Cremers. Depth super-resolution meets uncalibrated photometric stereo. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017. 2, 7

[46] Y. Quéau, J.-D. Durou, and J.-F. Aujol. Normal Integration: A Survey. *Journal of Mathematical Imaging and Vision*, 2017. 4

[47] Y. Quéau, J. Mélou, F. Castan, D. Cremers, and J.-D. Durou. A Variational Approach to Shape-from-shading Under Natural Illumination. In *Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR)*, 2017. 5

[48] R. Ramamoorthi and P. Hanrahan. An Efficient Representation for Irradiance Environment Maps. In *Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques*, pages 497–500, 2001. 3, 4

[49] S. R. Richter and S. Roth. Discriminative shape from shading in uncalibrated illumination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1128–1136, 2015. 3

[50] G. Riegler, M. Rüther, and H. Bischof. ATGV-net: accurate depth super-resolution. In *Proceedings of the European Conference on Computer Vision*, pages 268–284, 2016. 2

[51] E. Rouy and A. Tourin. A viscosity solutions approach to shape-from-shading. *SIAM Journal on Numerical Analysis*, 29(3):867–884, 1992. 3

[52] M. Schmidt. minFunc: unconstrained differentiable multivariate optimization in Matlab. `http://www.cs.ubc.ca/~schmidtm/Software/minFunc.html`, 2005. 5

[53] S. Schuon, C. Theobalt, J. Davis, and S. Thrun. Lidarboost: Depth superresolution for TOF 3D shape scanning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 343–350, 2009. 2

[54] E. Strekalovskiy and D. Cremers. Real-time minimization of the piecewise smooth Mumford-Shah functional. In *Proceedings of the European Conference on Computer Vision*, pages 127–141, 2014. 5

[55] M. Unger, T. Pock, M. Werlberger, and H. Bischof. A convex approach for variational super-resolution. In *DAGM Symposium*, pages 313–322, 2010. 2

[56] Y. Wang, W. Yin, and J. Zeng. Global convergence of ADMM in nonconvex nonsmooth optimization. *arXiv preprint arXiv:1511.06324*, 2015. 5

[57] C. Wu, M. Zollhöfer, M. Nießner, M. Stamminger, S. Izadi, and C. Theobalt. Real-time shading-based refinement for consumer depth cameras. *ACM Transactions on Graphics*, 33(6):200:1–200:10, 2014. 3

[58] J. Xie, R. S. Feris, and M.-T. Sun. Edge-guided single depth image super resolution. *IEEE Transactions on Image Processing*, 25(1):428–438, 2016. 2, 6, 7, 8

[59] J. Xie, R. S. Feris, S.-S. Yu, and M.-T. Sun. Joint super resolution and denoising from a single depth image. *IEEE Transactions on Multimedia*, 17(9):1525–1537, 2015. 2

[60] Q. Yang, R. Yang, J. Davis, and D. Nistér. Spatial-depth super resolution for range images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007. 2, 6, 7, 8

[61] L.-F. Yu, S.-K. Yeung, Y.-W. Tai, and S. Lin. Shading-based shape refinement of RGB-D images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1415–1422, 2013. 3

[62] R. Zhang, P.-S. Tsai, J. E. Cryer, and M. Shah. Shape-from-shading: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(8):690–706, 1999. 3

[63] M. Zollhöfer, A. Dai, M. Innman, C. Wu, M. Stamminger, C. Theobalt, and M. Nießner. Shading-based refinement on volumetric signed distance functions. *ACM Transactions on Graphics*, 34(4):96:1–96:14, 2015. 3

# Chapter 6

# Variational Uncalibrated Photometric Stereo under General Lighting

## Copyright


Bjoern Haefner, Zhenzhang Ye, Maolin Gao, Tao Wu, Yvain Quéau, and Daniel Cremers
**Variational Uncalibrated Photometric Stereo under General Lighting**
*2019 IEEE/CVF International Conference on Computer Vision (ICCV)*
DOI: 10.1109/ICCV.2019.00863

## Individual contributions

Significant contribution in realizing the scientific project.

| | |
|---|---|
| Problem definition | *significantly contributed* |
| Literature survey | *significantly contributed* |
| Implementation | *significantly contributed* |
| Experimental evaluation | *significantly contributed* |
| Preparation of the manuscript | *significantly contributed* |

In accordance with the *IEEE Thesis / Dissertation Reuse Permissions*, we include the accepted version of the original publication [6] in the following.

# Variational Uncalibrated Photometric Stereo under General Lighting

Bjoern Haefner*, [1] Zhenzhang Ye*, [1] Maolin Gao[2] Tao Wu[1] Yvain Quéau[3] Daniel Cremers[1]
[1]Technical University of Munich [2]Artisense [3]GREYC, UMR CNRS 6072
{bjoern.haefner, zz.ye, tao.wu, cremers}@tum.de maolin@artisense.ai yvain.queau@ensicaen.fr

## Abstract

*Photometric stereo (PS) techniques nowadays remain constrained to an ideal laboratory setup where modeling and calibration of lighting is amenable. To eliminate such restrictions, we propose an efficient principled variational approach to uncalibrated PS under general illumination. To this end, the Lambertian reflectance model is approximated through a spherical harmonic expansion, which preserves the spatial invariance of the lighting. The joint recovery of shape, reflectance and illumination is then formulated as a single variational problem. There the shape estimation is carried out directly in terms of the underlying perspective depth map, thus implicitly ensuring integrability and bypassing the need for a subsequent normal integration. To tackle the resulting nonconvex problem numerically, we undertake a two-phase procedure to initialize a balloon-like perspective depth map, followed by a "lagged" block coordinate descent scheme. The experiments validate efficiency and robustness of this approach. Across a variety of evaluations, we are able to reduce the mean angular error consistently by a factor of* 2–3 *compared to the state-of-the-art.*

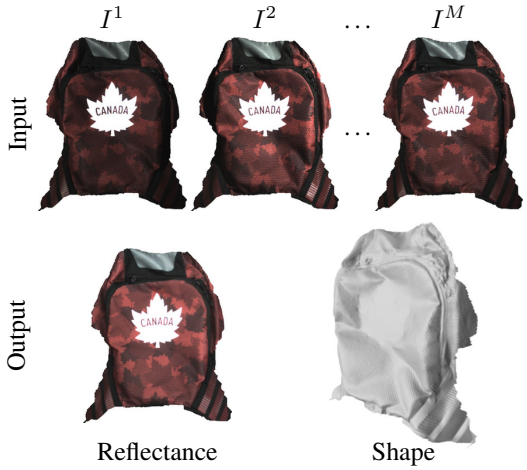



Figure 1. We present an efficient variational scheme to solve uncalibrated photometric stereo under general lighting. Given a set of input RGB images captured from the same viewing angle but under unknown, varying general illumination (top, $M = 20$ images were acquired in an office under daylight, while freely moving a hand-held LED light source), fine-detailed reflectance and shape (bottom, we show the estimated albedo and perspective depth maps) are recovered by an end-to-end variational approach.

## 1. Introduction

Photometric stereo techniques aim at acquiring both the shape and the reflectance of a scene. To this end, multiple images are acquired under the same viewing angle but varying lighting, and a physics-based image formation model is inverted. However, the classic way to solve this inverse problem requires lighting to be highly controlled, which restricts practical applications to laboratory setups where careful calibration of lighting must be carried out.

The objective of this research work is to simplify the overall photometric stereo pipeline, by providing an efficient solution to uncalibrated photometric stereo under general lighting, as illustrated in Figure 1 (the code is released[1]). In comparison with existing efforts in the same direction, the proposed one has the following advantages:

- The joint estimation of shape, reflectance and general lighting is formulated as an end-to-end, mathematically transparent variational problem;
- A real 3D-surface represented as a depth map is recovered, rather than possibly non-integrable normals;
- It is robust, due to the use of Cauchy's robust M-estimator and Huber-TV albedo regularization;
- It is computationally efficient, thanks to a tailored lagged block coordinate descent scheme initialized using a simple balloon-like shape.

After reviewing related works in Section 2, we discuss in Section 3 the image formation model considered in this work. It can be inverted using the variational approach in Section 4. A dedicated numerical solution is then introduced in Section 5 and empirically evaluated in Section 6. Section 7 eventually draws the conclusion of this research.

*Authors contributed equally.
[1]https://github.com/zhenzhangye/general_ups

## 2. Related Work

3D-models of scenes are essential in many applications such as visual inspection [14] or computer-aided surgery using augmented reality [12]. A 3D-model consists of geometric (position, orientation, etc.) and photometric (color, texture, etc.) properties. Given a set of photographies, the aim of 3D scanning is to invert the image formation process in order to recover these geometric and photometric properties of the observed scene. This notion thus includes both those of 3D-reconstruction (geometry) and of reflectance estimation (photometry).

Many approaches to the problem of 3D-reconstruction from photographies have been studied, and they are grouped under the generic naming "shape-from-X", where X stands for the clue which is being used (shadows [44], contours [10], texture [49], template [6], structured light [16], motion [35], focus [36], silhouettes [21], etc.). Geometric shape-from-X techniques are based on the identification and analysis of feature point or areas in the image. In contrast, photometric techniques build upon the analysis of the quantity of light received by each photosite of the camera's sensor. Among photometric techniques, *shape-from-shading* is probably the most famous one. This technique, developed in the 70s by Horn *et al.* [25], consists in 3D-reconstruction from a single image of a shaded scene. It is a classic ill-posed inverse problem whose numerical solving usually requires the surface's reflectance to be known [13]. In order both to limit the ambiguities of shape-from-shading and to allow for automatic reflectance estimation, it has been suggested to consider not just one image of the scene, but several ones acquired from the same viewing angle but under varying lighting. This variant, which was introduced in the late 70s by Woodham [50], is known as *photometric stereo*.

Among the various shape-from-X techniques mentioned above, photometric stereo is the only 3D-scanning technique i.e., the only one which is able to achieve both 3D-reconstruction and reflectance estimation. However, early photometric approaches strongly rely on the control of lighting. The latter is usually assumed for simplicity to be directional, although the case of nearby point light sources has recently regained some attention [31, 33]. More importantly, lighting is assumed to be calibrated. Indeed, the uncalibrated problem is ill-posed: the underlying normal map can be estimated only up to a linear ambiguity [20], which reduces to a generalized bas-relief one if integrability is enforced [9]. To resolve the latter ambiguity, some prior on the scene's surface or geometry must be introduced, see [48] for a recent survey. A natural way to enforce integrability consists in following a differential approach to photometric stereo [11, 32] i.e., directly estimate the 3D-surface as a depth map instead of first estimating the surface normals and then integrating them. Such a differential approach to photometric stereo can be coupled with variational methods in order to iteratively refine depth, reflectance and lighting in a robust manner [42]. In addition to the theoretical interest of enforcing integrability in order to limit ambiguities, differential approaches to photometric stereo have the advantages of easing combination with other 3D-reconstruction methods [17, 40], and of bypassing the problem of integrating the estimated normal field, which is by itself a non-trivial problem [41]. Besides, any error in the estimated normal field might propagate during integration, and thus robustness to specularities or shadows must be enforced during normal estimation, see again [48] for some discussion.

All the research works mentioned in the previous paragraph assume that lighting is induced by a single light source. Nevertheless, many studies rather considered the case of more general illumination conditions, which finds a natural application in outdoor conditions [43]. For instance, the apparent motion of the sun within a day induces changes in the illumination direction which, in theory, allow photometric stereo-based 3D-reconstruction. However, this apparent motion is close to being planar, and thus the rank of the set of illumination vectors is equal or close to 2 [45] (see also [23] for additional discussion on the stability of single-day photometric stereo). This situation is thus similar to the two-image case, which is known to be ill-posed since the early 90s [28, 37, 51], although it is still an active research area [29]. In order to limit the instabilities due to this issue, one possibility is to consider images acquired over many seasons as in [2, 3], or to resort to deep neural networks [22]. Another one is to consider a non-directional illumination model to represent natural illumination, as for instance in [26]. Modeling natural illumination is a promising track, since such a model would not be restricted to sunny days, and images acquired under cloudy days are known to yield more accurate 3D-reconstructions [23].

However, the previous approaches to photometric stereo under natural illumination assume calibrated lighting, where calibration is deduced from time and GPS coordinates or from a calibration target. The case of both general and uncalibrated lighting is much more challenging and has been fewly explored, apart from studies restricted to sparse 3D-reconstructions [46] or relying on the prior knowledge of a rough geometry [4, 27, 40, 47]. Uncalibrated photometric stereo under natural illumination has been revisited recently in [34], using a spatially-varying equivalent directional lighting model. However, results were limited to the recovery of possibly non-integrable surface normals. Instead, the method which we propose in the present paper directly recovers the underlying surface represented as a depth map. Following the seminal work of Basri and Jacobs [7], it considers the spherical harmonics representation of general lighting in lieu of the equivalent directional approximation, as discussed in the next section.

## 3. Image Formation Model

In photometric stereo (PS), we are given a number of observations $\{I^i\}_{i=1}^M$, each $I^i : \Omega \subset \mathbb{R}^2 \to \mathbb{R}^C$ representing a multi-channel image (i.e. $C \geq 1$) over a masked pixel domain $\Omega$. Assuming that the object being pictured is Lambertian, the surface's reflectance is represented by the albedo $\rho$, and the general image formation model is as follows, for all $i \in \{1, ..., M\}$, $c \in \{1, ..., C\}$, and $\mathbf{p} \in \Omega$:

$$I_c^i(\mathbf{p}) = \int_{\mathbb{S}^2} \rho_c(\mathbf{p}) \ell_c^i(\boldsymbol{\omega}) \max\{\boldsymbol{\omega} \cdot \mathbf{n}(\mathbf{p}), 0\} \, \mathrm{d}\boldsymbol{\omega}. \tag{1}$$

Here $\mathbb{S}^2$ is the unit sphere in $\mathbb{R}^3$, $\ell_c^i : \mathbb{S}^2 \to \mathbb{R}_+$ represents the channel-wise intensity of the incident light, and $\rho_c(\mathbf{p}) \in \mathbb{R}_+$ and $\mathbf{n}(\mathbf{p}) \in \mathbb{S}^2$ are the channel-wise albedos and the unit-length surface normals, respectively, at the surface point conjugate to pixel $\mathbf{p} \in \Omega$. The max operation in (1) encodes self-shadows. The overall integral $\int_{\mathbb{S}^2}$ collects elementary luminance contributions arising from all incident lighting directions $\boldsymbol{\omega}$. In the setup of uncalibrated PS, the quantities $\{\ell_c^i\}$, $\{\rho_c\}$, in addition to $\mathbf{n}$, are unknown.

*Equivalent directional lighting* [24] approximates (1) via

$$\begin{aligned} I_c^i(\mathbf{p}) &= \rho_c(\mathbf{p}) \, \bar{\ell}_c^i(\mathbf{p}) \cdot \mathbf{n}(\mathbf{p}), \\ \bar{\ell}_c^i(\mathbf{p}) &:= \int_{\{\boldsymbol{\omega} \in \mathbb{S}^2 \,:\, \boldsymbol{\omega} \cdot \mathbf{n}(\mathbf{p}) \geq 0\}} \ell_c^i(\boldsymbol{\omega}) \boldsymbol{\omega} \, \mathrm{d}\boldsymbol{\omega}. \end{aligned} \tag{2}$$

where $\bar{\ell}_c^i(\mathbf{p})$ represents the mean lighting over the visible hemisphere at $\mathbf{p}$. The field $\bar{\ell}_i^c$ is *spatially variant* but can be approximated by directional lighting over small local patches. Over each patch, one is thus faced with the ambiguities of directional uncalibrated PS [20]. State-of-the-art patch-wise methods [34] first solve this problem over each patch, then connect the patches to form a complete normal field up to rotation, and eventually estimate the rotation which best satisfies the integrability constraint. Errors may however get propagated during the sequence, resulting in a possibly non-integrable normal field.

Instead of such an equivalent directional lighting model, we rather consider a *spherical harmonic approximation* (SHA) of general lighting [8, 7]. By defining the half-cosine kernel $k$ as

$$k(\boldsymbol{\omega}, \mathbf{n}) := \max\{\boldsymbol{\omega} \cdot \mathbf{n}, 0\}, \tag{3}$$

we can view (1) as an analog of a convolution:

$$I_c^i(\mathbf{p}) = \rho_c(\mathbf{p}) \int_{\mathbb{S}^2} k(\boldsymbol{\omega}, \mathbf{n}(\mathbf{p})) \ell_c^i(\boldsymbol{\omega}) \, \mathrm{d}\boldsymbol{\omega}. \tag{4}$$

Invoking the Funk-Hecke theorem, we obtain the following harmonic expansion analogous to Fourier series:

$$\int_{\mathbb{S}^2} k(\boldsymbol{\omega}, \mathbf{n}(\mathbf{p})) \ell_c^i(\boldsymbol{\omega}) \, \mathrm{d}\boldsymbol{\omega} = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} (k_n \ell_{n,m}^{i,c}) h_{n,m}(\mathbf{n}(\mathbf{p})). \tag{5}$$

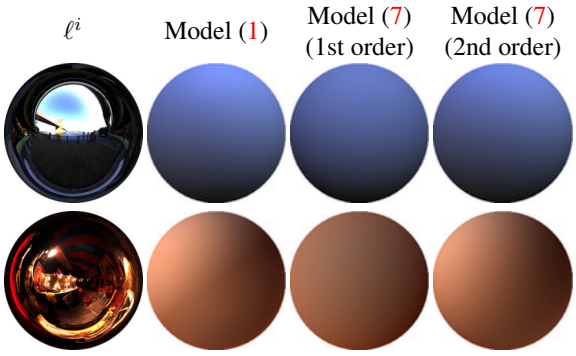



Figure 2. Illustration of RGB ($C = 3$) environment lighting $\ell^i = (\ell_1^i, \ell_2^i, \ell_3^i)$, the resulting images (assuming white albedos and a spherical shape) under the image formation model (1) and its approximation by spherical harmonics. The approximation by the second-order spherical harmonics is nearly perfect.

Here the spherical harmonics $\{h_{n,m}\}$ form an orthonormal basis of $L^2(\mathbb{S}^2)$, and $\{k_n\}$ and $\{\ell_{n,m}^{i,c}\}$ are the expansion coefficients of $k$ and $\ell_c^i$ with respect to $\{h_{n,m}\}$. Since most energy in the expansion (5) concentrates on low-order terms [8], we obtain the *second-order SHA* by truncating the series up to the first nine terms (i.e., $0 \leq n \leq 2$):

$$\int_{\mathbb{S}^2} k(\boldsymbol{\omega}, \mathbf{n}(\mathbf{p})) \ell_c^i(\boldsymbol{\omega}) \, \mathrm{d}\boldsymbol{\omega} \approx \sum_{n=0}^{2} \sum_{m=-n}^{n} (k_n \ell_{n,m}^{i,c}) h_{n,m}(\mathbf{n}(\mathbf{p})). \tag{6}$$

The first-order SHA refers to the truncation up to the first four terms (i.e., $0 \leq n \leq 1$). It is shown in [8] that, for distant lighting, at least 75% of the resulting irradiance is captured by the first-order SHA, and 98% by the second-order SHA (cf. Figure 2 for a visualization).

Plugging (6) and specifics of spherical harmonics [8] into (4), we finalize our image formation model as:

$$I_c^i(\mathbf{p}) \approx \rho_c(\mathbf{p}) \, \mathbf{l}_c^i \cdot \mathbf{h}[\mathbf{n}](\mathbf{p}), \tag{7}$$

$$\mathbf{h}[\mathbf{n}] = \left[1, \mathbf{n}_1, \mathbf{n}_2, \mathbf{n}_3, \mathbf{n}_1\mathbf{n}_2, \mathbf{n}_1\mathbf{n}_3, \mathbf{n}_2\mathbf{n}_3, \mathbf{n}_1^2 - \mathbf{n}_2^2, 3\mathbf{n}_3^2 - 1\right]^\top. \tag{8}$$

Here $\mathbf{h}[\mathbf{n}] : \Omega \to \mathbb{R}^9$ represents the second-order harmonic images, and $\mathbf{l}_c^i \in \mathbb{R}^9$ represents the harmonic lighting vector whose entries have absorbed $\{k_n \ell_{n,m}^{i,c}\}$ and constant factors of $\{h_{n,m}\}$. A key advantage of the SHA (7) over the equivalent directional lighting model (2) lies in the *spatial invariance* of the lighting vectors $\{\mathbf{l}_c^i\}$, which yields a less ill-posed inverse problem [7]. The counterpart is the nonlinear dependency upon the normal components, which we will handle in Section 5 using a tailored numerical solution. In the next section, we build upon the key observations that integrability [9] and perspective projection [39] both largely reduce the ambiguities of uncalibrated PS to derive a variational approach to inverting the SHA (7).

## 4. Variational Uncalibrated PS

In this section, we shall propose a joint variational model for uncalibrated PS. To this end, let a 3D-frame $(Oxyz)$ be attached to the camera, with $O$ the optical center, the $z$-axis aligned with the optical axis such that $z > 0$ for any 3D point $(x, y, z)$ in front of the camera. Further let a 2D-frame $(O'uv)$ be attached to the focal plane which is parallel to the $xy$-plane and contains the masked pixel domain $\Omega$. Under perspective projection, the surface geometry is modeled as a map $\mathbf{x} : \mathbf{p} = (u, v) \in \Omega \mapsto \mathbf{x}(u, v) \in \mathbb{R}^3$ given by

$$\mathbf{x}(u, v) = z(u, v)K^{-1}[u, v, 1]^\top, \tag{9}$$

with $z : \Omega \to \mathbb{R}_+$ the *depth* map and

$$K := \begin{bmatrix} f_u & 0 & u_0 \\ 0 & f_v & v_0 \\ 0 & 0 & 1 \end{bmatrix} \tag{10}$$

the calibrated camera's intrinsics matrix. In the following we denote for convenience $(\widetilde{u}, \widetilde{v}) := (u - u_0, v - v_0)$.

Assuming that $z$ is differentiable, the surface normal $\mathbf{n}$ at point $\mathbf{x}(u, v)$ is the unit vector oriented towards the camera such that $\mathbf{n}(u, v) \propto \partial_u \mathbf{x}(u, v) \times \partial_v \mathbf{x}(u, v)$, which yields the following parameterization of the normal by the depth:

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

$$\widetilde{\mathbf{n}}[z](u, v) := \begin{bmatrix} f_u \partial_u z(u, v) \\ f_v \partial_v z(u, v) \\ -z(u, v) - \widetilde{u}\, \partial_u z(u, v) - \widetilde{v}\, \partial_v z(u, v) \end{bmatrix}. \tag{12}$$

Note that the dependence of $\widetilde{\mathbf{n}}[z]$ on $z$ is linear.

Based on the forward model (7) and the parameterization (11) of normals, we formulate the joint recovery of reflectance, lighting and geometry as the following variational problem:

$$\min_{\{\rho_c\},\{\mathbf{l}_c^i\},z} \sum_{i=1}^{M} \sum_{c=1}^{C} \int_\Omega \phi_\lambda \Big( \rho_c(u, v)\, \mathbf{l}_c^i \cdot \mathbf{h}[\mathbf{n}[z]](u, v) - I_c^i(u, v) \Big) \,\mathrm{d}u\,\mathrm{d}v + \mu \sum_{c=1}^{C} \int_\Omega |\nabla \rho_c(u, v)|_\gamma \,\mathrm{d}u\,\mathrm{d}v. \tag{13}$$

In the first term above, we use *Cauchy's M-estimator* to penalize the data-fitting discrepancy:

$$\phi_\lambda(s) = \lambda^2 \log(1 + s^2/\lambda^2), \tag{14}$$

It is indeed well-known that Cauchy's estimator, being non-convex, is robust against outliers; see for instance [42] in the context of PS. The scaling parameter $\lambda = 0.15$ is used in all experiments.

The second term in (13) represents a Huber total-variation (TV) regularization on each albedo map $\rho_c$, with the Huber loss defined by

$$|s|_\gamma := \begin{cases} |s|^2/(2\gamma) & \text{if } |s| \le \gamma, \\ |s| - \gamma/2 & \text{if } |s| > \gamma, \end{cases} \tag{15}$$

and $\gamma = 0.1$ being fixed in the experiments. It turns out that the Huber TV imposes desirable smoothness on the albedo maps $\{\rho_c\}$ and in turn improves the joint estimation overall. Eventually, $\mu > 0$ is a weight parameter which balances the data-fitting term and the Huber TV one. Its value was empirically set to $2 \cdot 10^{-6}$ (see Section 6 for some discussion).

In (13), geometry is directly optimized in terms of the depth $z$ (rather than indirectly in terms of the normal $\mathbf{n}$). This both ensures integrability and avoids integration of normals into depths as a post-processing step.

## 5. Solver and Implementation

To solve the variational problem (13) numerically, we follow a "discretize-then-optimize" approach. There, $\Omega \subset \mathbb{R}^2$ is replaced by $\mathbb{R}^N$, $N$ being the number of pixels inside $\Omega$, which yields discretized vectors $z, \{\rho_c\}_{c=1}^C \in \mathbb{R}^N$. To alleviate notational burden, we sometimes refer to a pixel by its index $j \in \{1, \dots, N\}$ and sometimes by its position $\mathbf{p} = (u, v) \in \Omega$. The spatial gradient $\nabla$ is discretized using a forward difference stencil.

We shall apply a lagged block coordinate descent (LBCD) method to find a local minimum of the objective function in (23). Due to the (highly) non-convex nature of (23), initialization of optimization variables has a strong influence on the final solution. In our implementation, we initialize $\rho_{c,j} = \text{median}(\{I_{c,j}^i\}_{i=1}^M)$ for all $c, j$ and $\mathbf{l}_c^i = [0.2, 0, 0, -1, 0, 0, 0, 0, 0]^\top$ for all $c, i$. Moreover, during the first eight iterations we freeze the second-order spherical harmonics coefficients $(\mathbf{l}_c^i)_5 = (\mathbf{l}_c^i)_6 = \dots = (\mathbf{l}_c^i)_9 = 0$ i.e., we reconstruct using only first-order spherical harmonic approximation as a warm start. Most real-world scenes being convex, we initialize the depth $z$ as a balloon-like surface, as discussed in the following.

### 5.1. Depth Initialization

It is readily seen that a trivial constant initialization of the depth $z$ yields uniform vertically aligned normals $\mathbf{n}[z]$ and, hence, zero entries in the initial harmonic images $\mathbf{h}[\mathbf{n}[z]]$. This would cause non-meaningful updates on albedos $\{\rho_c\}$ and lighting vectors $\{\mathbf{l}_c^i\}$; cf. Figure 3 for an illustration.

To solve this issue, we specialize the depth initialization which undergoes two phases:

1. Following [38], we generate a balloon-like depth map $z_o$ under orthographic projection.

2. We then convert the orthographic depth $z_o$ to a perspective depth $z_p$ via normal integration [41].

Phase 1 is pursued via seeking a depth map $z_o$ which has minimal surface area subject to a constant volume $V$:

$$\begin{aligned} &\min_{z_o} \int_\Omega \sqrt{1+|\nabla z_o|^2}\,\mathrm{d}u\,\mathrm{d}v \\ &\text{s.t.} \int_\Omega z_o\,\mathrm{d}u\,\mathrm{d}v = V. \end{aligned} \tag{16}$$

A global minimizer of this model can be efficiently computed by simple projected gradient iterations:

$$z_o^{(k+1/2)} = z_o^{(k)} - \tau\nabla^\top\left(\frac{1}{\sqrt{1+|\nabla z_o^{(k)}|^2}}\nabla z_o^{(k)}\right), \tag{17}$$

$$z_o^{(k+1)} = z_o^{(k+1/2)} + \left(\frac{V - \int_\Omega z_o^{(k+1/2)}\,\mathrm{d}u\,\mathrm{d}v}{\int_\Omega \mathrm{d}u\,\mathrm{d}v}\right)\cdot \mathbf{1}_\Omega, \tag{18}$$

where $\mathbf{1}_\Omega(u,v)\equiv 1$ and $\tau = 0.8/\|\nabla\|_{\text{spec}}$ with $\|\cdot\|_{\text{spec}}$ the spectral norm. The volume constant $V$ is a hyperparameter which is empirically chosen, see Section 6 for discussion.

Next, we convert the orthographic depth $z_o$ to a perspective depth $z_p$. Note that $z_o$ complies with the orthographic projection, under which a 3D-point $\widehat{\mathbf{x}}$ is represented by

$$\widehat{\mathbf{x}}(u,v) = [u, v, z_o(u,v)]^\top, \tag{19}$$

and the corresponding surface normal $\widehat{\mathbf{n}}$ to the surface at $\widehat{\mathbf{x}}$ conjugate to pixel $\widehat{\mathbf{p}} = (u,v)$ is given by

$$\widehat{\mathbf{n}}(u,v) = \frac{1}{\sqrt{|\nabla z_o(u,v)|^2+1}}[\nabla z_o(u,v), -1]^\top. \tag{20}$$

Since $\widehat{\mathbf{n}}$ is invariant to the projection model, Eq. (11) also implies that

$$\widehat{\mathbf{n}}(u,v) \propto \begin{bmatrix} f_u\partial_u\widehat{z}_p(u,v) \\ f_v\partial_v\widehat{z}_p(u,v) \\ -1 - \widetilde{u}\partial_u\widehat{z}_p(u,v) - \widetilde{v}\partial_v\widehat{z}_p(u,v)\end{bmatrix}, \tag{21}$$

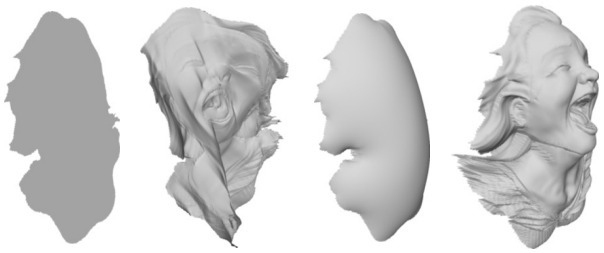
where $\widehat{z}_p(u,v) = \log z_p(u,v)$ stands for the log-perspective depth. This further implies the formula for $\nabla\widehat{z}_p$:

$$\nabla\widehat{z}_p(u,v) = \frac{-1}{\frac{\widetilde{u}\widehat{\mathbf{n}}_1(u,v)}{f_u} + \frac{\widetilde{v}\widehat{\mathbf{n}}_2(u,v)}{f_v} + \widehat{\mathbf{n}}_3(u,v)}\begin{bmatrix}\frac{1}{f_u}\widehat{\mathbf{n}}_1(u,v) \\ \frac{1}{f_v}\widehat{\mathbf{n}}_2(u,v)\end{bmatrix}, \tag{22}$$

which can be integrated to obtain $\widehat{z}_p$ (and hence $z_p$). The overall pipeline in Phase 2 is summarized as follows:

1. $(z_o \to \widehat{\mathbf{n}})$: Compute $\widehat{\mathbf{n}}$ by (20).
2. $(\widehat{\mathbf{n}} \to \nabla\widehat{z}_p)$: Compute $\nabla\widehat{z}_p$ by (22).
3. $(\nabla\widehat{z}_p \to z_p)$: Perform integration [41] to obtain $\widehat{z}_p$. Return $z_p = \exp\widehat{z}_p$ as the initialized (perspective) depth.

As discussed in [18] the perspective surface area depends linearly on the depth $z$. This complicates direct perspective ballooning, since the depth is driven towards zero and hence yields numerical instability. For this reason, we opted for the two-step approach which bypasses the issue.



Trivial $z_p \equiv 1$ and its result $z$ Our $z_p$ and its result $z$

Figure 3. Impact of depth initialization: a trivial constant initialization on the left vs. our initialization on the right and its corresponding resulting geometry estimates. Further results from varying initializations can be found in the supplementary material.

### 5.2. Lagged Block Coordinate Descent

Even with a reasonable initialization, the numerical resolution of Problem (23) remains challenging. Due to the appearances of the spherical harmonic approximation $\mathbf{h}[\mathbf{n}[z]]$ and the Cauchy's M-estimator $\phi_\lambda$, the objective in (23) is highly nonlinear and nonconvex. To tackle these challenges, here we present a lagged block coordinate descent (LBCD) method which performs efficiently in practice.

To derive LBCD, we introduce an auxiliary variable $\theta \in \mathbb{R}^N$ such that $\theta_j = |\widetilde{\mathbf{n}}_j[z]|$. This enables us to rewrite (11) as $\mathbf{n}_j[z] = \widetilde{\mathbf{n}}_j[z]/\theta_j$. Then we formulate the following constrained optimization problem:

$$\begin{aligned} \min_{\theta,\{\rho_c\},\{\mathbf{l}_c^i\},z} &\sum_{i=1}^{M}\sum_{c=1}^{C}\sum_{j=1}^{N}\phi_\lambda\left(r_{i,c,j}(\theta_j,\rho_{c,j},\mathbf{l}_c^i,z)\right) \\ &+ \mu\sum_{c=1}^{C}\sum_{j=1}^{N}|(\nabla\rho_c)_j|_\gamma, \\ \text{s.t. } &\theta_j = |\widetilde{\mathbf{n}}_j[z]|,\ \forall j\in\{1,\dots,N\}, \end{aligned} \tag{23}$$

where $r_{i,c,j}$ is the residual function defined by:

$$r_{i,c,j}(\theta_j,\rho_{c,j},\mathbf{l}_c^i,z) = \rho_{c,j}\,\mathbf{l}_c^i\cdot\mathbf{h}_j[\widetilde{\mathbf{n}}_j[z]/\theta_j] - I_{c,j}^i. \tag{24}$$

Upon initialization, the proposed LBCD proceeds as follows. At iteration $k$, we lag $\theta$ one iteration behind, i.e.,

$$\theta_j^{(k+1)} := |\widetilde{\mathbf{n}}_j[z^{(k)}]|, \quad \forall j\in\{1,\dots,N\}, \tag{25}$$

and then sequentially update each of the three blocks (namely $\{\rho_c\}$, $\{\mathbf{l}_c^i\}$ and $z$). In each resulting subproblem,

we solve (lagged) weighted least squares problems as an approximation of the Cauchy loss and/or the Huber loss. This is detailed in the following:

- (Update $\{\rho_c\}$): We evaluate the residual

$$r_{i,c,j}^{(k+1/3)} := r_{i,c,j}(\theta_j^{(k+1)}, \rho_{c,j}^{(k)}, \mathbf{l}_c^{i,(k)}, z^{(k)}), \tag{26}$$

and then set up the (lagged) weight factors for both the Cauchy loss and the Huber loss as

$$w_{i,c,j}^{(k+1/3)} := \phi'_\lambda(r_{i,c,j}^{(k+1/3)})/r_{i,c,j}^{(k+1/3)}, \tag{27}$$

$$q_{c,j}^{(k+1/3)} := 1/\max\{\gamma, |(\nabla\rho_c^{(k)})_j|\}. \tag{28}$$

The albedos $\{\rho_c\}$ are updated as the solution to the following linear weighted least-squares problem:

$$\{\rho_c^{(k+1)}\} := \arg\min_{\{\rho_c\}} \mu \sum_{c,j} q_{c,j}^{(k+1/3)} |(\nabla\rho_c)_j|^2 + \sum_{i,c,j} w_{i,c,j}^{(k+1/3)} |r_{i,c,j}(\theta_j^{(k+1)}, \rho_j^c, \mathbf{l}_c^{i,(k)}, z^{(k)})|^2, \tag{29}$$

which is carried out by conjugate gradient (CG).

- (Update $\{\mathbf{l}_c^i\}$): The lighting subproblem is similar to the one for albedos, except for absence of the Huber TV term. Upon evaluation of the residual $r_{i,c,j}^{(k+2/3)}$ and the weight factor $w_{i,c,j}^{(k+2/3)}$, we update $\{\mathbf{l}_c^i\}$ by solving the following linear weighted least-squares problem via CG:

$$\{\mathbf{l}_c^{i,(k+1)}\} = \arg\min_{\mathbf{l}_c^i} \sum_{i,c,j} w_{i,c,j}^{(k+2/3)} \cdot |r_{i,c,j}(\theta_j^{(k+1)}, \rho_{c,j}^{(k+1)}, \mathbf{l}_c^i, z^{(k)})|^2. \tag{30}$$

- (Update $z$): The depth subproblem requires additional efforts. With $r_{i,c,j}^{(k+1)}$ and $w_{i,c,j}^{(k+1)}$ evaluated after the $\{\mathbf{l}_c^i\}$-update, we are faced with the following weighted least squares problem:

$$\min_z \sum_{i,c,j} w_{i,c,j}^{(k+1)} |r_{i,c,j}(\theta_j^{(k+1)}, \rho_{c,j}^{(k+1)}, \mathbf{l}_c^{i,(k+1)}, z)|^2, \tag{31}$$

where the dependence of $r_{i,c,j}$ on $z$ is still nonlinear. Therefore, we further linearize $r_{i,c,j}$ with respect to $z$ and arrive at the following update:

$$z^{(k+1)} = \arg\min_z \sum_{i,c,j} w_{i,c,j}^{(k+1)} \cdot |r_{i,c,j}^{(k+1)} + J_r(z^{(k)})(z - z^{(k)})|^2, \tag{32}$$

where $J_r(z^{(k)})$ is the Jacobian of the map $z \mapsto r_{i,c,j}(\theta_j^{(k+1)}, \rho_{c,j}^{(k+1)}, \mathbf{l}_c^{i,(k+1)}, z)$ at $z = z^{(k)}$. The resulting linearized least-squares problem is again solved by CG. In our experiments, we additionally incorporate backtracking line search in the $z$-update to ensure a monotonic decrease of the energy.

# 6. Experimental Validation

This section is concerned with the evaluation of the proposed nonconvex variational approach to uncalibrated photometric stereo under general lighting.

## 6.1. Synthetic Experiments

To validate the impact of the initial volume $V$ in (16), the tunable hyper-parameter $\mu$, and the number of input images $M$ in (13), we consider 36 challenging synthetic datasets. We use four different depth maps ("Joyful Yell" [1], "Lucy" [30], "Armadillo" [30] and "Thai Statue" [30]) and nine different albedo maps and each of those 36 combinations is rendered as described in (1) using $M = 25$ different environment maps[2], cf. Figure 4. The resulting 25 RGB images per dataset are used as input, along with the intrinsic camera parameters and a binary mask $\Omega$. A quantitative evaluation on the triplet $(V, \mu, M)$ is carried out on four randomly chosen datasets (Armadillo & White albedo, Joyful Yell & Ebsd albedo, Lucy & Hippie albedo, and Thai Statue & Voronoi albedo), comparing the impact of each value of $(V, \mu, M)$ on the resulting mean angular error (MAE) between ground truth and estimated normals.

First, we validate the choice of the input volume $V$ using the initially fixed values of $\mu = 2 \cdot 10^{-6}$ and $M = 25$. As the volume depends on the size of the mask, we consider a linear parametrization $V(\kappa) = \kappa|\Omega| = \kappa N$ and evaluate a range of ratios $\kappa \in [1, 10^3]$. Figure 5 (left) indicates that the optimal value of $\kappa$ is dataset-dependent. For synthetic datasets we always selected this optimal value, yet for real-world data no such evaluation is possible and $\kappa$ must be tuned manually. Since the ballooning-based depth initialization can be carried out in real-time (implementation is parallelized in CUDA), the user has an immediate feedback on the initial depth and thus a plausible initial shape is easily drawn. Humans excel at estimating size and shape of objects [5] and real-world experiments will show that a manual choice of $\kappa$ can result in appealing geometries.

Next, we evaluate the impact of $\mu$, cf. Figure 5 (right). As can be seen, the depth estimate seems to deteriorate for too small and too large values of $\mu$, whereas $\mu \in [10^{-6}, 10^{-5}]$ seems to provide good depth estimates across all albedo maps. Therefore we fix $\mu = 2 \cdot 10^{-6}$ for all our upcoming experimental evaluation.

Unsurprisingly, the MAE is inversely proportional to the number $M$ of input images, but runtime increases (linearly) with $M$, cf. Figure 6. We found that $M \in [15, 25]$ represents a good trade-off between runtime and accuracy, and fix $M = 20$ for all our further experiments. Our Matlab implementation needs about 1–2 minutes on a computer with an Intel $i7$ processor.

[2] Environment maps are downloaded from http://www.hdrlabs.com/sibl/archive.html

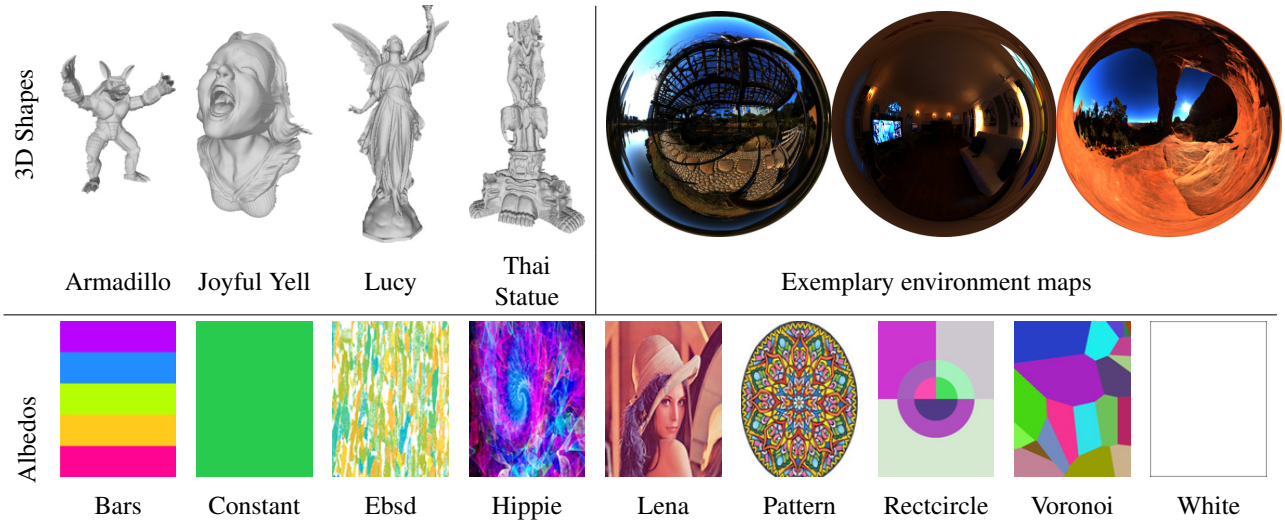



Figure 4. The four 3D-shapes and nine albedo maps we used to create 36 (3D-shape, albedo) datasets. For each dataset, $M = 25$ images were rendered using different environment maps such as those shown on the top right.

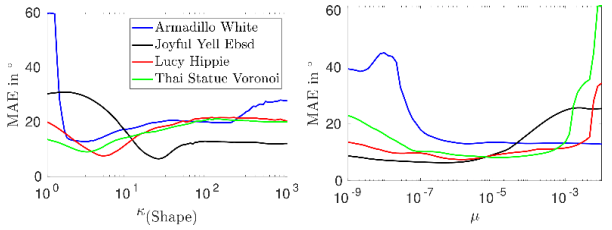



Figure 5. Impact of the initial volume $V_{(\text{Shape})}$ as well as $\mu$ on the accuracy of the estimated depth. Based on these experiments we choose $\kappa_{(\text{Armadillo})} = 2.84$, $\kappa_{(\text{Joyful Yell})} = 24.77$, $\kappa_{(\text{Lucy})} = 4.98$, $\kappa_{(\text{Thai Statue})} = 3.05$ and $\mu = 2 \cdot 10^{-6}$ for all experiments, where $V_{(\text{Shape})} = \kappa_{(\text{Shape})} N_{(\text{Shape})}$.

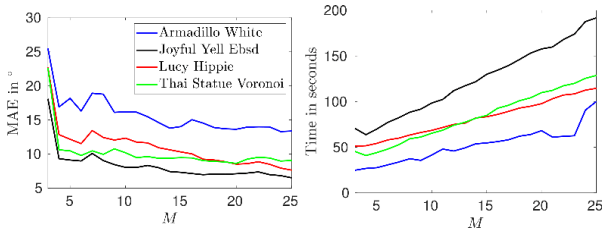



Figure 6. Impact of the number of images $M$ on the mean angular error (MAE) and the runtime. Based on these insights we choose $M = 20$ for our experiments.

Having fixed the choice of $(V, \mu, M)$, we can now evaluate our approach against other state-of-the-art methods. We compare our results against those obtained by an uncalibrated photometric stereo approach assuming directional lighting [15], and another one assuming general (first-order spherical harmonics) illumination yet relying on an input shape prior (e.g., from an RGB-D sensor) [40]. As this limiting assumption on the access to a sensor-based depth prior is not always given and to make comparison fair, we input as depth prior to this method the ballooning initialization described in Section 5.1. Furthermore, we compare against another uncalibrated photometric stereo work under natural illumination [34][3], which resorts to the equivalent directional lighting instead of spherical harmonics, cf. Section 3. Table 1 shows the median and mean MAEs over all 36 datasets (a more detailed table can be found in the supplementary material). On these datasets, it can be seen that our method quantitatively outperforms the current state-of-the-art by a factor of 2–3. This gain is also evaluated qualitatively in Figure 7, which shows a selection of two results.

| Approach | [15] | [40] | [34] | Ours |
|---|---|---|---|---|
| Median | 27.16 | 21.14 | 34.06 | **9.17** |
| Mean | 34.15 | 21.18 | 35.53 | **10.72** |

Table 1. Median and mean of the mean angular errors (MAE) over all 36 datasets. The proposed approach overcomes the state-of-the-art by a factor of 2–3.

### 6.2. Real-World Experiments

For real-world data we use the publicly available dataset of [19]. It offers eight challenging real-world datasets of objects with complex geometry and albedo captured under daylight and a freely moving LED, along with intrinsics matrix $K$ and masks $\Omega$. Results are presented in Figure 8. Despite relying on a directional lighting model, the approach of [15] produces reasonable results on some datasets (Face1, Ovenmitt or Shirt), but it fails on others. As [40] assumes a reliable prior on depth in order to perform a photometric refinement, this approach is biased towards its initialization and thus, only when the depth prior is very close to

[3] Code associated with [15] and [40] can be found online, and the results obtained by [34] were provided by the authors.

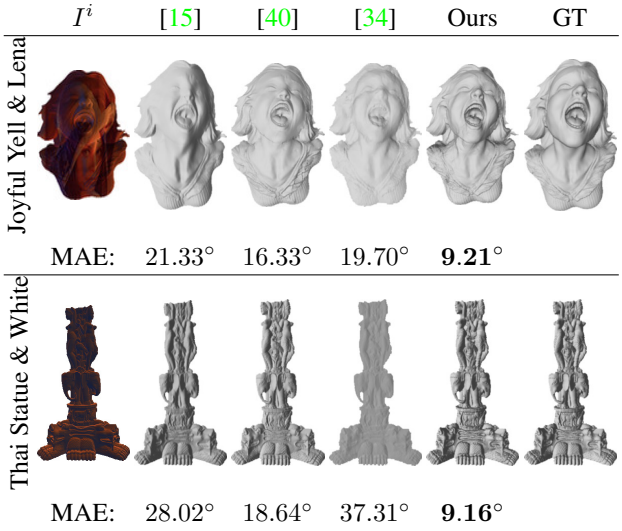



Figure 7. Results of state-of-the-art approaches and our approach on two out of the 36 synthetic datasets. Numbers show the mean angular error (MAE) in degrees.

the objects' rough shape (Ovenmitt, Shirt, Tabletcase, Vase) a meaningful geometry is recovered. The approach of [34] estimates a possibly non-integrable normal field only, and it can be seen that after integration the depth map might not be satisfactory. As our approach optimizes over depth directly, such issues are not apparent and we are able to recover fine-scale geometric details throughout all tests.

## 7. Conclusion

We proposed a variational approach to uncalibrated photometric stereo (PS) under general lighting. Assuming a perspective camera setup, our method jointly estimates shape, reflectance and lighting in a robust manner. The possible non-integrability of normals is bypassed by the direct estimation of the underlying depth map, and robustness is ensured by resorting to Cauchy's M-estimator and Huber-TV albedo regularization. Although the problem is nonconvex and thus numerically challenging and initialization-dependent, we tackled it efficiently through a tailored lagged block coordinate descent algorithm and ballooning-based depth initialization. Over a series of evaluations on synthetic and real data, we demonstrated that our method outperforms existing methods in terms of MAE by a factor of 2–3 and provides highly detailed reconstructions even in challenging real-world settings.

In future research, a more automated balloon-like depth initialization is desirable. Exploring the theoretical foundations (uniqueness of a solution) of differential perspective uncalibrated PS under spherical harmonic lighting and analyzing the convergence properties of the proposed numerical scheme constitute two other promising perspectives.

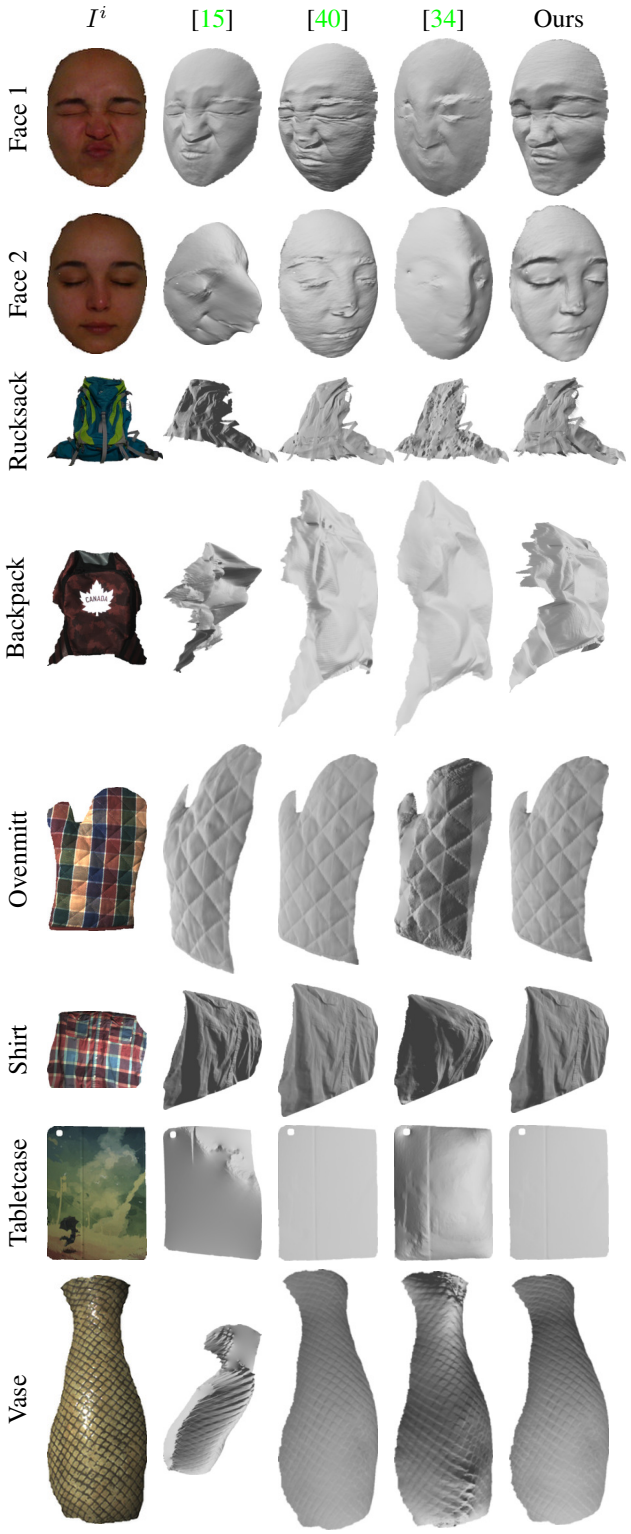



Figure 8. Results of state-of-the-art approaches and our approach on challenging real-world datasets. While the competing approaches fail on some datasets, our approach consistently yields satisfactory results.

## References


[1] The Joyful Yell. 2015. https://www.thingiverse.com/thing:897412. 6

[2] Austin Abrams, Christopher Hawley, and Robert Pless. Heliometric Stereo: Shape from Sun Position. In *European Conference on Computer Vision (ECCV)*, volume 7573 of *Lecture Notes in Computer Science*, pages 357–370, 2012. 2

[3] Jens Ackermann, Fabian Langguth, Simon Fuhrmann, and Michael Goesele. Photometric stereo for outdoor webcams. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 262–269, 2012. 2

[4] Jens Ackermann, Martin Ritz, André Stork, and Michael Goesele. Removing the example from example-based photometric stereo. In *Trends and Topics in Computer Vision (ECCV Workshops)*, volume 6554 of *Lecture Notes in Computer Science*, pages 197–210. 2012. 2

[5] Joseph Baldwin, Alistair Burleigh, Robert Pepperell, and Nicole Ruta. The perceived size and shape of objects in peripheral vision. *i-Perception*, 7(4):2041669516661900, 2016. 6

[6] Adrien Bartoli, Yan Gérard, Francois Chadebecq, Toby Collins, and Daniel Pizarro. Shape-from-template. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(10):2099–2118, 2015. 2

[7] Ronen Basri, David Jacobs, and Ira Kemelmacher. Photometric stereo with general, unknown lighting. *International Journal of Computer Vision*, 72(3):239–257, 2007. 2, 3

[8] Ronen Basri and David W Jacobs. Lambertian reflectances and linear subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(2):218–233, 2003. 3

[9] Peter N Belhumeur, David J Kriegman, and Alan L Yuille. The bas-relief ambiguity. *International Journal of Computer Vision*, 35(1):33–44, 1999. 2, 3

[10] Michael Brady and Alan Yuille. An extremum principle for shape from contour. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(3):288–301, 1984. 2

[11] Manmohan Chandraker, Jiamin Bai, and Ravi Ramamoorthi. On Differential Photometric Reconstruction for Unknown, Isotropic BRDFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2941–2955, Dec 2013. 2

[12] Toby Collins and Adrien Bartoli. 3D Reconstruction in Laparoscopy with Close-Range Photometric Stereo. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 7511 of *Lecture Notes in Computer Science*, pages 634–642. 2012. 2

[13] Jean-Denis Durou, Maurizio Falcone, and Manuela Sagona. Numerical Methods for Shape-from-shading: A New Survey with Benchmarks. *Computer Vision and Image Understanding*, 109(1):22–43, 2008. 2

[14] Abdul Rehman Farooq, Melvyn Lionel Smith, Lyndon Neal Smith, and Sagar Midha. Dynamic photometric stereo for on line quality control of ceramic tiles. *Computers in industry*, 56(8-9):918–934, 2005. 2

[15] Paolo Favaro and Thoma Papadhimitri. A closed-form solution to uncalibrated photometric stereo via diffuse maxima. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 821–828, 2012. 7, 8

[16] Jason Geng. Structured-light 3D surface imaging: a tutorial. *Advances in Optics and Photonics*, 3(2):128–160, 2011. 2

[17] Paulo FU Gotardo, Tomas Simon, Yaser Sheikh, and Iain Matthews. Photogeometric scene flow for high-detail dynamic 3d reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 846–854, 2015. 2

[18] Gottfried Graber, Jonathan Balzer, Stefano Soatto, and Thomas Pock. Efficient minimal-surface regularization of perspective depth maps in variational stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 511–520, 2015. 5

[19] Bjoern Haefner, Songyou Peng, Alok Verma, Yvain Quéau, and Daniel Cremers. Photometric depth super-resolution. *Arxiv preprint 1809.10097*, 2018. 7

[20] Hideki Hayakawa. Photometric stereo under a light source with arbitrary motion. *Journal of the Optical Society of America A*, 11(11):3079–3089, 1994. 2, 3

[21] Carlos Hernández. *Stereo and Silhouette Fusion for 3D Object Modeling from Uncalibrated Images Under Circular Motion*. Thèse de doctorat, École Nationale Supérieure des Télécommunications, 2004. 2

[22] Yannick Hold-Geoffroy, Paulo FU Gotardo, and Jean-François Lalonde. Deep photometric stereo on a sunny day. *Arxiv preprint 1803.10850*, 2018. 2

[23] Yannick Hold-Geoffroy, Jinsong Zhang, Paulo FU Gotardo, and Jean-François Lalonde. What is a good day for outdoor photometric stereo? In *International Conference on Computational Photography*, 2015. 2

[24] Yannick Hold-Geoffroy, Jinsong Zhang, Paulo FU Gotardo, and Jean-François Lalonde. $x$-hour outdoor photometric stereo. In *International Conference on 3D Vision*, 2015. 3

[25] Berthold KP Horn. *Shape From Shading: A Method for Obtaining the Shape of a Smooth Opaque Object From One View*. PhD thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 1970. 2

[26] Jiyoung Jung, Joon-Young Lee, and In So Kweon. One-Day Outdoor Photometric Stereo Using Skylight Estimation. *International Journal of Computer Vision*, 2019. (to appear). 2

[27] Ira Kemelmacher-Shlizerman and Ronen Basri. 3d face reconstruction from a single image using a single reference face shape. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(2):394–405, 2010. 2

[28] Ryszard Kozera. On Shape Recovery from Two Shading Patterns. *International Journal of Pattern Recognition and Artificial Intelligence*, 6(4):673–698, 1993. 2

[29] Ryszard Kozera and Alexander Prokopenya. Second-order algebraic surfaces and two image photometric stereo. In *International Conference on Computer Vision and Graphics*, pages 234–247, 2018. 2

[30] Marc Levoy, J Gerth, B Curless, and K Pull. The stanford 3d scanning repository. 2005. http://www-graphics.stanford.edu/data/3dscanrep. 6

[31] Fotios Logothetis, Roberto Mecca, and Roberto Cipolla. Semi-calibrated near field photometric stereo. In *Proceed-*

*ings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 941–950, 2017. 2

[32] Roberto Mecca, Ariel Tankus, Aaron Wetzler, and Alfred M Bruckstein. A direct differential approach to photometric stereo with perspective viewing. *SIAM Journal on Imaging Sciences*, 7(2):579–612, 2014. 2

[33] Roberto Mecca, Aaron Wetzler, Alfred M Bruckstein, and Ron Kimmel. Near field photometric stereo with point light sources. *SIAM Journal on Imaging Sciences*, 7(4):2732–2770, 2014. 2

[34] Zhipeng Mo, Boxin Shi, Feng Lu, Sai-Kit Yeung, and Yasuyuki Matsushita. Uncalibrated photometric stereo under natural illumination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2936–2945, 2018. 2, 3, 7, 8

[35] Theo Moons, Luc Van Gool, and Maarten Vergauwen. 3D Reconstruction from Multiple Images. *Foundations and Trends in Computer Graphics and Vision*, 4(4):287–404, 2008. 2

[36] SK Nayar Y Nakagawa and SK Nayar. Shape from focus. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(8):824–831, 1994. 2

[37] Ruth Onn and Alfred Bruckstein. Integrability disambiguates surface recovery in two-image photometric stereo. *International Journal of Computer Vision*, 5(1):105–113, 1990. 2

[38] Martin R Oswald, Eno Toeppe, and Daniel Cremers. Fast and Globally Optimal Single View Reconstruction of Curved Objects. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 534–541, 2012. 4

[39] Thoma Papadhimitri and Paolo Favaro. A new perspective on uncalibrated photometric stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1474–1481, 2013. 3

[40] Songyou Peng, Bjoern Haefner, Yvain Quéau, and Daniel Cremers. Depth super-resolution meets uncalibrated photometric stereo. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, pages 2961–2968, 2017. 2, 7, 8

[41] Yvain Quéau, Jean-Denis Durou, and Jean-François Aujol. Normal Integration: A Survey. *Journal of Mathematical Imaging and Vision*, 60(4):576–593, 2018. 2, 4, 5

[42] Yvain Quéau, Tao Wu, François Lauze, Jean-Denis Durou, and Daniel Cremers. A Non-Convex Variational Approach to Photometric Stereo under Inaccurate Lighting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 350–359, 2017. 2, 4

[43] Yoichi Sato and Katsushi Ikeuchi. Reflectance analysis under solar illumination. In *Workshop on Physics-Based Modeling in Computer Vision (ICCV Workshops)*, pages 180–187, 1995. 2

[44] Steven A Shafer and Takeo Kanade. Using shadows in finding surface orientations. *Computer Vision, Graphics, and Image Processing*, 22(1):145–176, 1983. 2

[45] Fangyang Shen, Kalyan Sunkavalli, Nicolas Bonneel, Szymon Rusinkiewicz, Hanspeter Pfister, and Xin Tong. Time-lapse photometric stereo and applications. *Computer Graphics Forum*, 33(7):359–367, 2014. 2

[46] Li Shen and Ping Tan. Photometric stereo and weather estimation using internet images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1850–1857, 2009. 2

[47] Boxin Shi, Kenji Inose, Yasuyuki Matsushita, Ping Tan, Sai-Kit Yeung, and Katsushi Ikeuchi. Photometric stereo using internet images. In *International Conference on 3D Vision*, volume 1, pages 361–368, 2014. 2

[48] Boxin Shi, Zhe Wu, Zhipeng Mo, Dinglong Duan, Sai-Kit Yeung, and Ping 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. 2

[49] Andrew P Witkin. Recovering surface shape and orientation from texture. *Artificial Intelligence*, 17(1):17–45, 1981. 2

[50] Robert J Woodham. Reflectance Map Techniques for Analyzing Surface Defects in Metal Castings. Technical Report MIT AITR-457, 1978. 2

[51] Jun Yang, Noboru Ohnishi, and Noboru Sugie. Two image photometric stereo method. In *Intelligent Robots and Computer Vision XI: Biological, Neural Net, and 3D Methods*, volume 1826 of *Proceedings of the International Society for Optical Engineering*, pages 452–463, 1992. 2

# Chapter 7

# Photometric Segmentation: Simultaneous Photometric Stereo and Masking

## Copyright


Bjoern Haefner, Yvain Quéau, and Daniel Cremers
**Photometric Segmentation: Simultaneous Photometric Stereo and Masking**
*2019 International Conference on 3D Vision (3DV)*
DOI: 10.1109/3DV.2019.00033

## Individual contributions

Leading role in realizing the scientific project.

| | |
|---|---|
| Problem definition | *significantly contributed* |
| Literature survey | *significantly contributed* |
| Implementation | *significantly contributed* |
| Experimental evaluation | *significantly contributed* |
| Preparation of the manuscript | *significantly contributed* |

In accordance with the *IEEE Thesis / Dissertation Reuse Permissions*, we include the accepted version of the original publication [4] in the following.

# Photometric Segmentation: Simultaneous Photometric Stereo and Masking

Bjoern Haefner
TU Munich
Munich, Germany
bjoern.haefner@tum.de

Yvain Quéau
GREYC, UMR CNRS 6072
Caen, France
yvain.queau@ensicaen.fr

Daniel Cremers
TU Munich
Munich, Germany
cremers@tum.de

## Abstract

*This work is concerned with both the 3D-reconstruction of an object using photometric stereo, and its 2D-segmentation from the background. In contrast with previous works on photometric stereo which assume that a mask of the area of interest has been computed beforehand, we formulate 3D-reconstruction and 2D-segmentation as a joint problem. The proposed variational solution combines a differential formulation of photometric stereo with the classic Chan-Vese model for active contours. Given a set of photometric stereo images, this solution simultaneously infers a binary mask of the object of interest and a depth map representing its 3D-shape. Experiments on real-world datasets confirm the soundness of simultaneously solving both these classic computer vision problems, as the joint approach considerably simplifies the overall 3D-scanning process for the end-user.*

## 1. Introduction

Photometric stereo [26] can be employed to estimate the 3D-shape of an object, given a set of images taken under the same viewing angle, but varying illumination. To this end, an image formation model describing the interactions between light and matter is inverted. Recent advances in the field have focused on relaxing several assumptions such as those of Lambertian reflectance and of calibrated lighting [23], in order to make the technique applicable in real-world scenarios and to simplify the 3D-scanning process.

However, all the approaches to photometric stereo which have been proposed so far assume that the area of interest is known in advance (see Figure 1). This means that the end-user is still required to perform a pre-segmentation of the object to reconstruct, before the 3D-reconstruction can be carried out. Such a pre-segmentation can be tedious and time-consuming: it would be way more convenient to achieve it automatically, while taking into account the information conveyed by the multi-light acquisition process of photometric stereo.

(a) (b) (c)

Figure 1. Given a set of input images such as (a), and a mask of the object to reconstruct (b), photometric stereo techniques infer 3D-geometry, represented in (c) under the form of surface normals. The present work aims at simplifying this process, by automatically achieving 2D-segmentation of the object in the same time as its 3D-reconstruction. That is to say, we aim at directly recovering the mask (b) and geometry (c) from the images (a).

In this work, we introduce a way to simultaneously achieve the 2D-segmentation of the object and its 3D-reconstruction, instead of first masking the object and then estimating its geometry. Building upon both the celebrated active contours model of Chan and Vese for two-region segmentation [2], and a recent PDE-based variational formulation of photometric stereo [20], we propose a joint variational approach to this problem. It comes down to estimating a minimal-length curve separating background from an area where the image formation model is satisfied and thus shape estimation is possible.

The proposed variational formulation for simultaneous 3D-reconstruction by photometric stereo and 2D-segmentation is detailed in Section 3, after discussing related variational approaches to photometric stereo and segmentation in Section 2. The resulting optimisation problem is numerically challenging, but recasting it as a level-set problem allows one to use classic convex optimisation techniques. Experimental results presented in Section 4 demonstrate the potential of this joint approach. Eventually, Section 5 concludes this study and suggests future research directions.

## 2. Variational Methods for Photometric Stereo and Segmentation

Assuming Lambertian reflectance with known directional lighting, neglecting shadows and assuming the object to reconstruct is pre-segmented, the classic formulation of photometric stereo [26] with $m$ images consists in solving a set of equations such as

$$I_i(x) = \rho(x)\,\mathbf{n}(x)\cdot\mathbf{s}_i, \quad \forall x\in\Omega, i\in\{1,\dots,m\}, \quad (1)$$

with $\Omega\subset\mathbb{R}^2$ the mask of the object to reconstruct, $I_i:\Omega\to\mathbb{R}$ the $i$-th input graylevel image, $\rho$ the reflectance (albedo) map, $\mathbf{n}$ the normal map (which encodes the 3D-geometry), and $\mathbf{s}_i\in\mathbb{R}^3$ a vector representing the incident lighting in the $i$-th image (in intensity and direction). Most of recent works on photometric stereo have focused on relaxing the assumptions of Lambertian reflectance (i.e., handling surfaces which exhibit a specular behavior) [4, 21, 11, 29] and calibrated directional lighting (i.e., handling unknown or non-uniform lighting) [5, 10, 13, 22], see for instance [23] for some discussion and [3] for a state-of-the-art joint solution to both issues using deep neural networks. However, in all of these recent works the object to reconstruct is assumed to be segmented a priori: the whole pipeline relies on the knowledge of the domain $\Omega$.

In order to get rid of this assumption, we will make use of the recent variational formulation of photometric stereo exposed in [20], and also advocated in [6, 10, 12, 24]. Therein, 3D-reconstruction by photometric stereo is formulated as a variational problem aiming at directly reconstructing the underlying depth map $z:\Omega\to\mathbb{R}$, thus bypassing the need for normal estimation followed by normal integration. It is an optimisation-based approach of the form

$$\min_z \int_\Omega \mathcal{P}_{\text{PS}}(z(x))\,\mathrm{d}x, \quad (2)$$

with $\mathcal{P}_{\text{PS}}(z(x))$ a term evaluating the pixel-wise discrepency between the data and a differential formulation of the image formation model (1). Under orthographic projection, the normal is linked to the gradient $\nabla z:\Omega\to\mathbb{R}^2$ of the underlying depth map $z:\Omega\to\mathbb{R}$ according to $\mathbf{n}(x)=\frac{[\nabla z(x),-1]^\top}{\sqrt{|\nabla z(x)|^2+1}}$ [19], thus from a pair of equations such as (1), with $i\neq j$, one gets, for any $x\in\Omega$:

$$\frac{I_i(x)}{[\nabla z(x),-1]^\top\cdot\mathbf{s}_i}=\frac{\rho(x)}{\sqrt{|\nabla z(x)|^2+1}}=\frac{I_j(x)}{[\nabla z(x),-1]^\top\cdot\mathbf{s}_j}, \quad (3)$$

from which one can deduce:

$$\mathbf{a}_{ij}(x)\nabla z(x)=b_{ij}(x), \quad (4)$$

with $\mathbf{a}_{ij}(x):=\begin{bmatrix}I_i(x)s_j^1-I_j(x)s_i^1\\ I_i(x)s_j^2-I_j(x)s_i^2\end{bmatrix}^\top\in\mathbb{R}^{1\times 2}$, $b_{ij}(x):=I_i(x)s_j^3-I_j(x)s_i^3\in\mathbb{R}$ and where $\mathbf{s}_i=\left[s_i^1,s_i^2,s_i^3\right]^\top$. This gives rise to $\binom{m}{2}$ different linear PDEs in $z$, which can be combined in a variational framework. Adding an arbitrary depth prior $z_0$ for numerical stability, depth estimation can then be formulated as (2), with

$$\mathcal{P}_{\text{PS}}(z(x)):=\frac{1}{\binom{m}{2}}\sum_{ij}\left(\mathbf{a}_{ij}(x)\nabla z(x)-b_{ij}(x)\right)^2+\lambda(z(x)-z_0(x))^2, \quad (5)$$

where $\lambda>0$ is some hyper-parameter.

On the other hand, there is a large amount of literature on the image segmentation problem. Let us mention for instance early approaches based on region merging heuristics [17], active contours evolving towards edges in the images (aka snakes) [8], or recent deep learning frameworks [1]. Another class of methods is based on piecewise-smooth approximation of the input image [14], which comes down to image segmentation in the case of piewise-constant approximation. A classic example of such an approach is the Chan-Vese active contour model [2]. It aims at estimating a minimal-length curve $C$ separating the image domain $\Omega$ between an area inside $C$ where the graylevel image $I$ is well-approximated by some value $\mu_1$, and an area outside $C$ where it is better-approximated by $\mu_2$. This can be formulated as follows:

$$\min_{\mu_1,\mu_2,C}\int_{\text{inside}(C)}\mathcal{P}_1(\mu_1,I(x))\,\mathrm{d}x+\int_{\text{outside}(C)}\mathcal{P}_2(\mu_2,I(x))\,\mathrm{d}x+\nu\,\text{length}(C), \quad (6)$$

where $\nu\geq 0$ is a hyper-parameter controlling the length of the curve $C$, and $\mathcal{P}_j(\mu_j,I(x))=(\mu_j-I(x))^2$, $j\in\{1,2\}$, such that the values $\mu_1,\mu_2$ resemble the mean intensity of the image $I$ in the region inside and outside $C$, respectively.

Image segmentation and 3D-reconstruction may appear as two disconnected problems. Nevertheless, each task contributes information which may be interesting for the other, and the joint solving of these inverse problems has proven valuable, for instance, in the context of dense multi-view reconstruction [7], X-ray tomography [9], pose estimation [18], SLAM [25] or hyperspectral imaging [28]. Inspired by such joint approaches to simultaneous reconstruction and segmentation, in the rest of this work we revisit the photometric stereo problem in the case where no prior segmentation of the object has been performed i.e., domain $\Omega$ in (1) is unknown. This is detailed in the next section, where we propose a joint approach to photometric stereo and masking which combines the variational photometric stereo formulation (2) with the Chan-Vese variational segmentation approach (6).

## 3. Photometric Segmentation

The underlying assumption of the classic Chan-Vese segmentation model is that the brightness in the foreground largely differs from that of the background: (6) assumes that brightness in the foreground and background are well-approximated by two different constants $\mu_1$ and $\mu_2$. In the context of photometric stereo images, the objects are usually captured in the dark, thus it would make sense to assume that the background has a constant, low, brightness. Yet, the foreground may contain shadowed or low-albedo areas which might wrongly be classified as background.

The redundancy of information induced by the lighting variations contributes useful information to overcome shadowing issues. One could for instance apply a similar method as (6), but on the stack of $m$ photometric stereo images. This can be simply achieved by setting $\mathcal{P}_j(\mu_j, I(x)) = \frac{1}{m}\sum_{i=1}^{m}(\mu_j - I_i(x))^2$ in (6). However, Figure 2 shows that this first approach remains unsatisfactory as it cannot distinguish between low albedo and background.

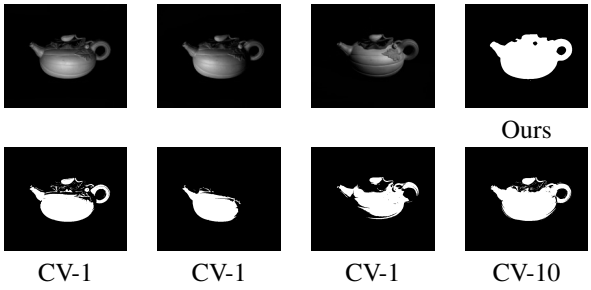



Figure 2. Top: three photometric stereo images (out of $m = 10$), and the segmentation obtained using the proposed joint reconstruction and segmentation method. Bottom: result of Chan-Vese segmentation applied to the single photometric stereo image shown above (CV-1), or to the whole set of 10 images (CV-10). It is difficult to segment a single image due to the ambiguity between background and shadows. Using multiple images improves results, but not as much as incorporating a photometric stereo model into segmentation, as we propose.

Instead of such a naive adaptation of the Chan-Vese model based on the average brightness, we suggest to drive segmentation by the image formation model. We define foreground as the set of pixels where the depth map $z$ can be estimated from the differential photometric stereo model (4). On the contrary, we define background as the set of pixels where the photometric stereo model is not valid i.e., any uniform depth map $z_0$ can be set in (4). This way, the resulting depth map $z$ will exhibit discontinuities along the separating curve $C$, which shall not fit to model (4). That is to say, it would "cost" more to wrongly include the boundaries of the object in the background or foreground, than to fit the curve separating foreground and background on the genuine object's boundaries.

In variational terms, this comes down to solving the following optimisation problem:

$$\begin{aligned}\min_{z,C} &\int_{\text{inside}(C)} \mathcal{P}_{\text{PS}}(z(x))\,\mathrm{d}x \\ &+ \int_{\text{outside}(C)} \mathcal{P}_{\text{PS}}(z_0(x))\,\mathrm{d}x \\ &+ \nu\,\text{length}(C),\end{aligned} \tag{7}$$

where $\mathcal{P}_{\text{PS}}$ is the differential photometric stereo fitting criterion defined in (5), $\nu \geq 0$ is a hyper-parameter controlling the length of the segmenting curve, and $z_0$ is an arbitrary depth prior which can be set using e.g., a prior on the camera-scene distance, or any arbitrary value if depth estimation up to an additive offset is acceptable (e.g., in our experiments we use $z_0 \equiv 1$).

Let us remark that the proposed variational model (7) differs from the Chan-Vese model (6) in two ways. First, the unknown depth $z$ needs not being estimated in the background, which slightly simplifies the process. On the other hand the unknown depth is spatially-varying on the foreground, yet the estimation of such a spatially-varying quantity is made possible by the multiple image measurements under varying lighting.

Despite these differences, the optimisation problem (7) contains the same major difficulty as the origin problem of Chan and Vese: it involves both 2D (the depth map) and 1D (the curve) entities, which makes optimisation nontrivial. Besides, optimisation over the curve would require an appropriate parameterisation, which is known to yield nontrivial numerical issues such as setting the number of control points and uniformly sampling them over the curve. In the rest of this section we introduce an equivalent level-sets [16] formulation of the problem, which yields a simpler numerical solution. That is to say, we embed the problem in a higher-dimensional space where it is easier to solve numerically.

Let us define the curve $C$ as the zero level-set of some function $\phi : \Omega \to \mathbb{R}$, such that $\phi \geq 0$ defines foreground and $\phi < 0$ defines background. Let us further denote by $H$ the Heaviside step function ($H(x) = 1$ if $x \geq 0$ and $H(x) = 0$ elsewhere). Then, (7) is rewritten as follows:

$$\begin{aligned}\min_{z,\phi} &\int_{\Omega} H(\phi(x))\mathcal{P}_{\text{PS}}(z(x))\,\mathrm{d}x \\ &+ \int_{\Omega} (1 - H(\phi(x)))\mathcal{P}_{\text{PS}}(z_0(x))\,\mathrm{d}x \\ &+ \nu \int_{\Omega} |\nabla H(\phi(x))|\,\mathrm{d}x,\end{aligned} \tag{8}$$

where optimisation is now carried out over two real-valued 2D maps over $\Omega$, and the segmenting curve $C$ can be computed a posteriori from $\phi$ by thresholding.

To jointly solve the problems of photometric stereo ($z$-estimation) and segmentation ($\phi$-estimation), we solve (8) alternatingly over each variable. At iteration $(k)$, we solve:

$$z^{(k+1)} = \arg\min_{z} \int_{\Omega} H(\phi^{(k)}(x)) \mathcal{P}_{\text{PS}}(z(x))\, \mathrm{d}x, \tag{9}$$

$$\begin{aligned} \phi^{(k+1)} = \arg\min_{\phi} & \int_{\Omega} H(\phi(x)) \mathcal{P}_{\text{PS}}(z^{(k+1)}(x))\, \mathrm{d}x \\ & + \int_{\Omega} (1 - H(\phi(x))) \mathcal{P}_{\text{PS}}(z_0(x))\, \mathrm{d}x \\ & + \nu \int_{\Omega} |\nabla H(\phi(x))|\, \mathrm{d}x. \end{aligned} \tag{10}$$

Problem (9) is a linear least squares problem, which can be solved using conjugate gradient iterations on the normal equations. Problem (10) is solved using gradient descent on the Euler-Lagrange equation

$$\begin{aligned} \delta(\phi(x)) \Bigg[ & \mathcal{P}_{\text{PS}}(z^{(k+1)}(x)) - \mathcal{P}_{\text{PS}}(z_0(x)) \\ & - \nu \operatorname{div}\left( \frac{\nabla \phi(x)}{|\nabla \phi(x)|} \right) \Bigg] = 0, \end{aligned} \tag{11}$$

where $\delta(\phi(x))$ is a dirac delta, which can be considered as the derivative of the heaviside function $H(\phi(x))$.

Before the first iteration we initialise $z^{(0)}$ with the prior $z_0$, while the initial foreground is a circle of radius 10: $\phi^{(0)} = 10 - \sqrt{(x_1 - c_1)^2 + (x_2 - c_1)^2}$, where $x_i$ are pixel coordinates and $c_i$ corresponds to the center of the image, $i \in \{1, 2\}$. We stop iterations when the relative energy between two consecutive iterations falls under a threshold of 0.02, and we noticed that this was achieved in at most 20 iterations.

## 4. Experimental Validation

This section provides experimental results for the proposed variational photometric segmentation model. Real-world datasets of 10 photometric stereo images (see Figure 3) are extracted from a publicly available challenging photometric stereo benchmark [23] (images were preprocessed using [27] in order to fit the Lambertian assumption). The ten images are chosen such that the object is illuminated from every direction. All the experiments were conducted using Matlab on a standard laptop with 16GB of RAM and an Intel Core $i7$ with 2.2GHz. Convergence was always reached in at most 1 minute.

The impact of the hyper-parameter $\nu$ on the segmentation result will first be discussed. This is followed by a quantitative and qualitative evaluation of our segmentation results against other segmentation approaches. Eventually, we show that the estimated geometry of the scene is on average better, and en par compared to the result using no mask, and the ground-truth mask, respectively.

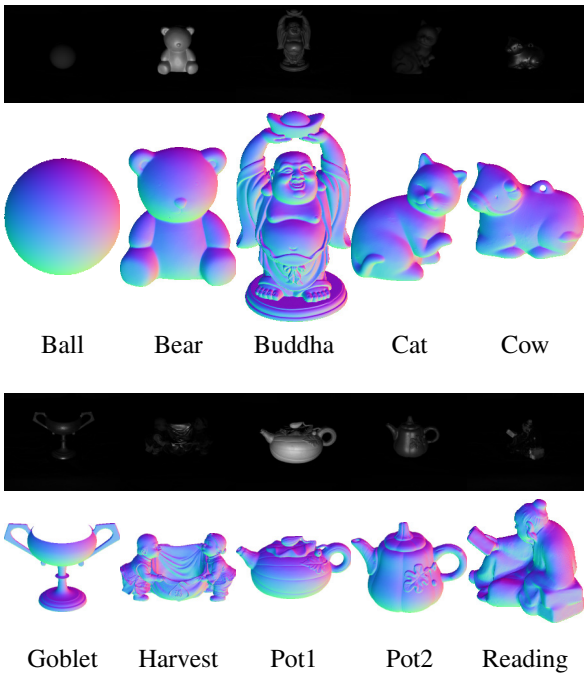
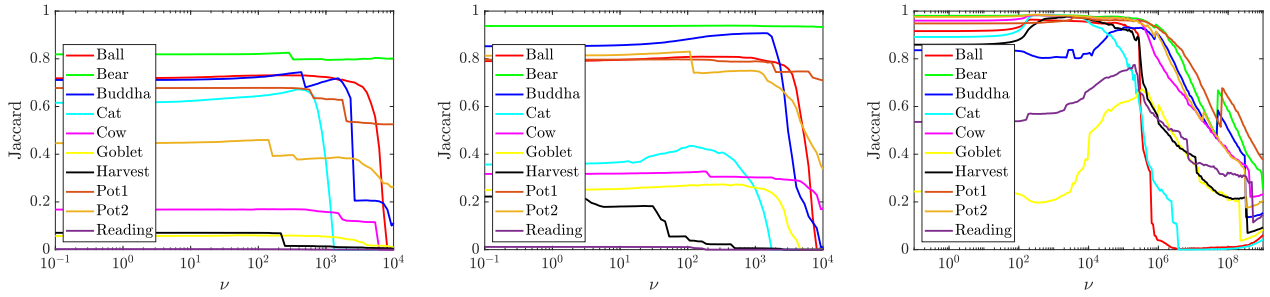



Figure 3. One out of ten grayscale images from [23] we used as input, along with the corresponding ground-truth normals.

### 4.1. Parameter Tuning

Our model comprises two tuning parameters $\lambda$ and $\nu$. In all our experiments we set $\lambda$ to a very small value of $10^{-9}$, since regularisation inside the mask is only intended to fix the translation ambiguity $z(x) := z(x) + \text{constant}$ in (4). Indeed, any small value of $\lambda$ will solve this ambiguity and ensure convexity of the optimisation problem with respect to $z$, yet a high value of $\lambda$ might bias the solution towards $z_0$. The tuning parameter $\nu$ is more crucial, hence we are going to evaluate it more thoroughly.

To this end we run our algorithm on the publicly available dataset [23] for values of $\nu \in [10^{-1}, 10^{9}]$ and evaluate the segmentation result using the Jaccard coefficient $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$, where $A$ and $B$ are two sets, in our case the ground truth mask and the estimated one. Our approach is compared against the classic Chan-Vese approach [2] presented in (6), which depends on the parameter $\nu$ as well. To show the impact of the choice of images in [2] we deploy two schemes. The first scheme uses a single image (randomly chosen from the data set) and we denote this approach with CV-1. The second scheme uses the same 10 images we use in the proposed method and we denote this approach with CV-10. The impact of the hyper-parameter $\nu$ can be evaluated in Figure 4. Not surprisingly, the Chan-Vese approach with ten images performs on average better than CV-1, as more data is used.




(a) CV-1 (b) CV-10 (c) Ours

Figure 4. Impact of the tuning parameter $\nu$ on the Jaccard coefficient (closer to one is better) for the classic Chan-Vese model (6) based on a single image (a) or multiple images (b), and for the proposed model (c). Our method systematically overcomes the Chan-Vese ones, provided that $\mu$ is not set too high in order to avoid over-segmentation.

Our model overcomes both Chan-Vese results throughout the evaluated range of $\nu$, which shows that a tailored photometric segmentation cost function helps to find a better estimate of the mask of the object. In the rest of the evaluation we use the value of $\nu$ which provides the best results, for both Chan-Vese methods and for the proposed one.

### 4.2. Segmentation Accuracy

To quantitatively validate the accuracy of the estimated mask we compare the Jaccard coefficients against those obtained by three different segmentation approaches. The first one is GIMP's "Foreground Select Tool" which is based on statistical models of color variation [15] and can be considered as a standard way to generate a mask for an end-user of photometric stereo. It asks him to draw two scribbles in the image which provide the best statistical information in terms of color to separate background from foreground. The two Chan-Vese approaches (with a single or multiple images) already discussed in the previous paragraph are also considered, along with our approach with the best possible parameters for each dataset based on our evaluation in Figure 4. All quantitative and qualitative results can be seen in Table 1 and Figure 5. As already shown in Figure 4, the proposed approach overcomes both CV-1 and CV-10, which proves that the photometric term improves segmentation results. GIMP performs best on the Goblet dataset since it is able to separate the inner part of it, which no other method is able to do. Still, GIMP, CV-1 and CV-10 mainly suffer from oversegmentation, as they are not intended to distinguish background from shading. Only Harvest shows an undersegmented result, where GIMP considers too much background as foreground, due to too much dark shading variations in the object. Although the automated GIMP tool seems to overcome classic Chan-Vese segmentation, it can not keep up with our joint approach, which delivers the best segmentation results overall.

| Dataset | GIMP | CV-1 | CV-10 | Proposed |
|---|---|---|---|---|
| Ball | 0.6958 | 0.7307 | 0.8090 | **0.9643** |
| Bear | 0.8767 | 0.8254 | 0.9391 | **0.9827** |
| Buddha | 0.9112 | 0.7441 | 0.9074 | **0.9320** |
| Cat | 0.8567 | 0.6719 | 0.4352 | **0.9842** |
| Cow | 0.5536 | 0.1695 | 0.3277 | **0.9829** |
| Goblet | **0.8706** | 0.0601 | 0.2734 | 0.6727 |
| Harvest | 0.4830 | 0.0706 | 0.2227 | **0.9773** |
| Pot1 | 0.7930 | 0.6781 | 0.7978 | **0.9727** |
| Pot2 | 0.9145 | 0.4596 | 0.8305 | **0.9851** |
| Reading | 0.4970 | 0.0023 | 0.0114 | **0.7748** |

Table 1. Quantitative evaluation of the segmentation obtained using GIMP, CV-1, CV-10 and the proposed method, based on the Jaccard coefficient. The proposed approach overcomes the others in most cases.

### 4.3. Normal Reconstruction Accuracy

We also question whether an accurate segmentation may improve the quality of the estimated geometry. The best possible estimate is obtained using photometric stereo with ground truth mask i.e., solving (8) with fixed $\phi$. Hence we consider the latter estimate as our "ground truth" geometry (as Figures 3 and 6 illustrate, such a baseline geometry may deviate from the real ground truth one, yet using the latter would bias the evaluation). For quantitative evaluation we calculate the mean angular error (MAE) in degrees between the baseline normals (resulting from the ground truth mask) and the estimated ones. As estimated normals we consider two approaches. The first one is an approach without any mask, that is every pixel is considered as valuable data point and used during optimisation. The second one is the proposed photometric segmentation approach. To make comparison fair, for both approaches we only evaluate MAE in the area corresponding to the intersection between the mask estimated by our approach and the ground truth one. Results can be seen in Table 2 and Figure 6.

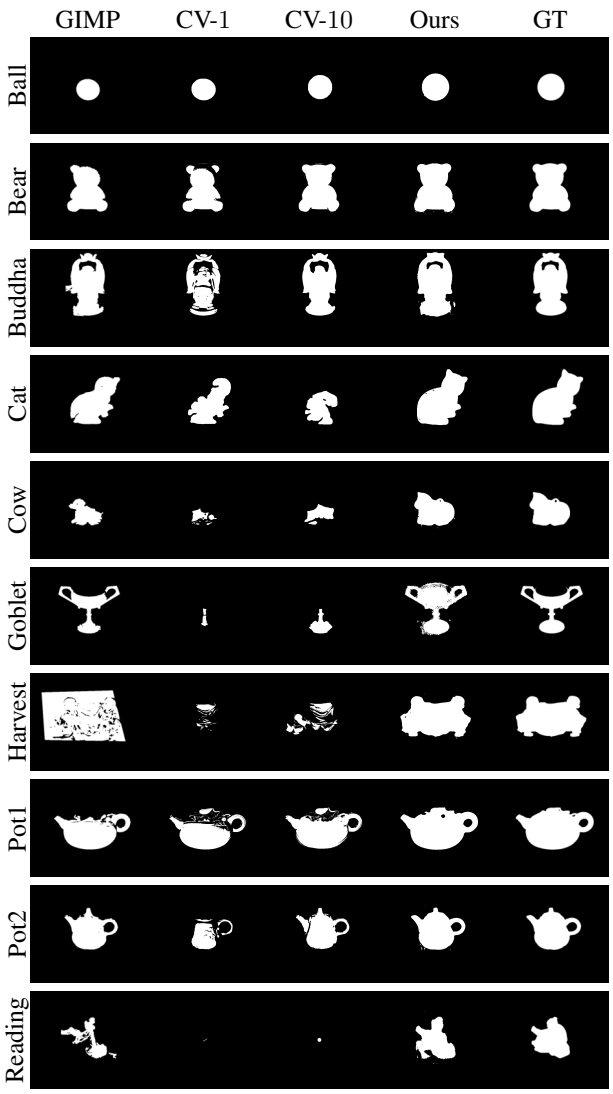



Figure 5. Qualitative comparison between the segmentation obtained using GIMP, CV-1, CV-10 and the proposed method which jointly estimates mask and geometry. In all the examples except Goblet, the estimated mask is the most accurate one.

These results show that geometry estimation indeed benefits from a joint 3D-reconstruction and segmentation approach: in most datasets our approach deviates much less from the baseline normals, compared to the approach without mask. Only the two data sets Harvest and Reading appear to perform better with no mask, but the loss in accuracy ($0.02^\circ$ and $0.15^\circ$, respectively) can be considered negligible. We believe that this gain comes from the fact that when using no mask, geometry is smoothed at the boundaries of the object, while when using a mask (or, when automatically finding this mask, as we propose) much sharper geometry can be recoverd near the boundaries.

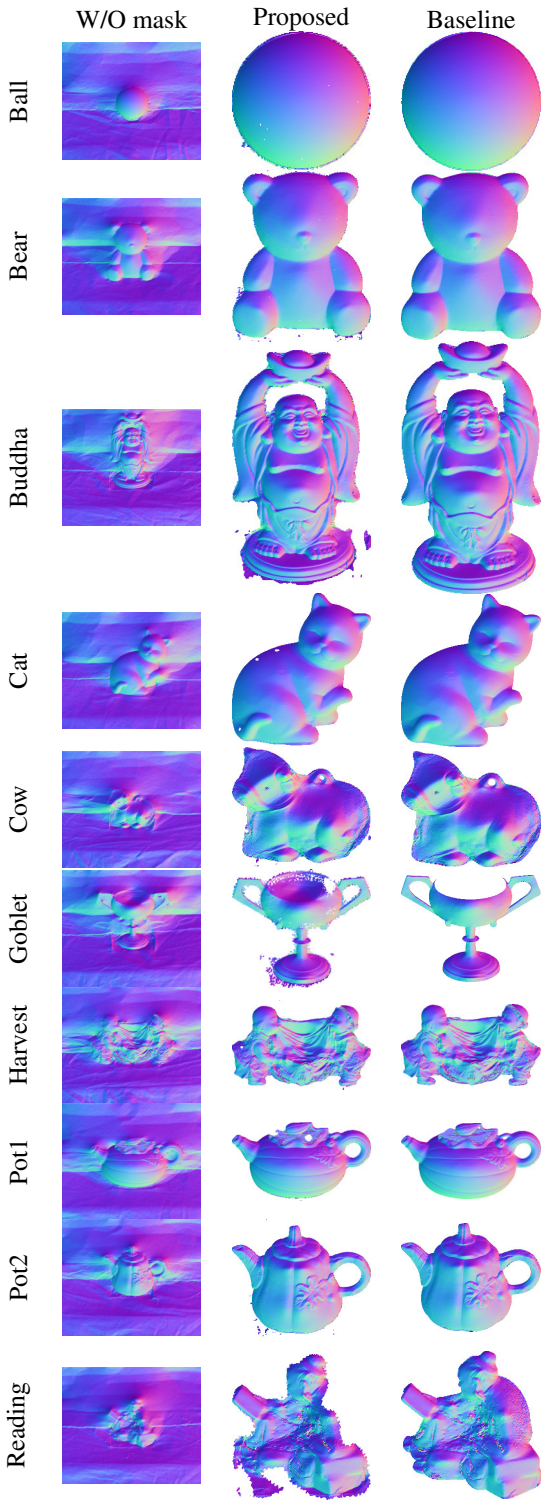



Figure 6. Qualititative results of the normal estimates using no mask and the proposed approach which jointly estimates mask and geometry. Estimated masks can be seen in Figure 5.

| Dataset | W/O mask | Proposed |
|---|---|---|
| Ball | 0.9290 | **0.7522** |
| Bear | 0.6211 | **0.2974** |
| Buddha | 0.7791 | **0.5370** |
| Cat | 0.2068 | **0.0868** |
| Cow | 0.9644 | **0.6592** |
| Goblet | 6.8144 | **6.4709** |
| Harvest | **0.6204** | 0.6816 |
| Pot1 | 1.7623 | **1.5196** |
| Pot2 | 0.8353 | **0.3747** |
| Reading | **9.0507** | 9.2291 |

Table 2. Comparison of the mean angular error (in degrees) on the estimated normals with respect to the baseline. The proposed approach slightly improves the geometry estimate in most cases.

Indeed, Figure 7 shows that the improvement becomes apparent at the boundaries of the estimates, where our approach has less error and larger deviation from the approach with no mask is visible. Especially in the case of the Buddha, where a large error appears in the hole of the arms and head, our approach has much less error, as it is able to detect this region as background. Only the results of Goblet and Reading largely deviate from the baseline normals inside the object. The difficulty with Goblet is the discontinuity in the upper part, which our approach is not able to detect. This results in geometry estimates across the discontinuity, inducing smoothing which deteriorates the overall geometry inside the object. Reading itself is very dark compared to the other objects, cf Figure 3. This causes the mask estimate to suffer from undersegmentation of the object.

## 5. Conclusion

We presented a joint variational approach to photometric stereo and segmentation. To the best of our knowledge this is the first methodology providing a scheme which is able to perform photometric stereo without the need of a mask. The proposed approach simplifies the photometric stereo process from the end-user perspective, by circumventing the need for tedious masking of the object and providing an end-to-end framework for object 3D-reconstruction, as shown in Figure 8. Experiments conducted on real-world benchmarks provided empirical evidence for the superiority of model-driven segmentation over naive segmentation based on brightness. Still, the proposed alternating optimisation strategy could be accelerated a lot by relying on parallel computing: in the future this will enable real-time results which will ease the setting of the hyper-parameter controlling the length of the boundary curve.

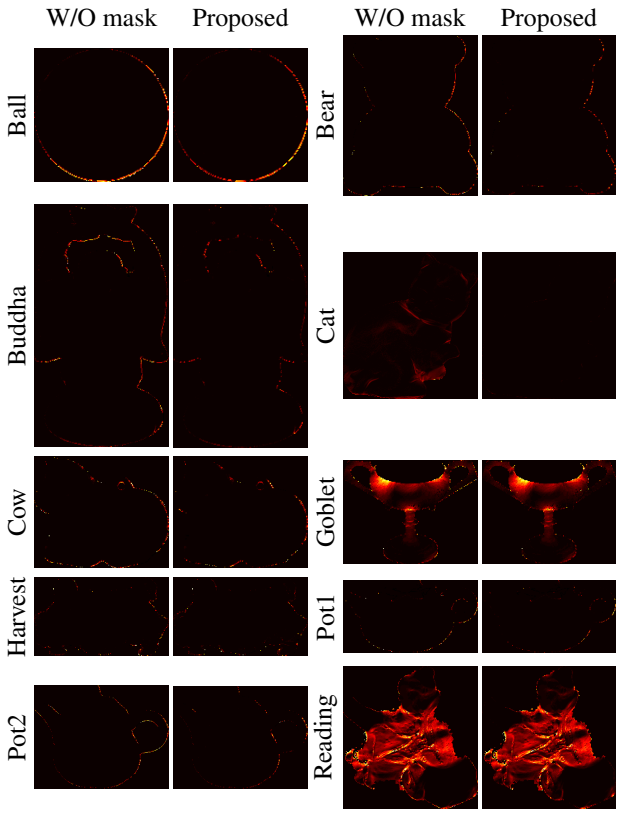



Figure 7. Spatial distribution of the angular error with respect to the baseline. Joint reconstruction and segmentation slightly reduces errors near the boundaries of the objects, since geometry estimate does not propagate over possibly discontinuous regions. Best seen in color on the Buddha and Cat examples.

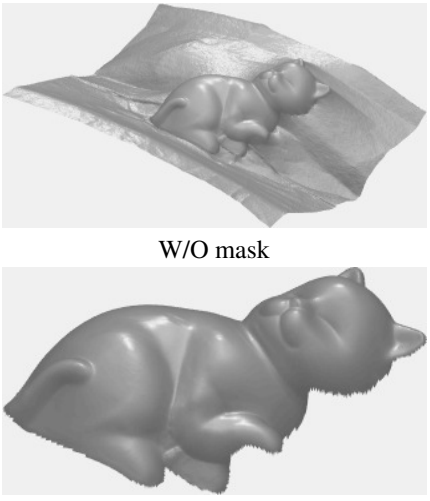



Figure 8. 3D-reconstruction of the Cat without masking, or using the proposed method. The latter makes possible an end-to-end 3D-scanning pipeline of objects.

## References


[1] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017. 2

[2] T. F. Chan and L. A. Vese. Active contours without edges. *IEEE Transactions on Image Processing*, 10(2):266–277, 2001. 1, 2, 4

[3] G. Chen, K. Han, B. Shi, Y. Matsushita, and K.-Y. K. Wong. Self-calibrating Deep Photometric Stereo Networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[4] K. H. M. Cheng and A. Kumar. Revisiting Outlier Rejection Approach for Non-Lambertian Photometric Stereo. *IEEE Transactions on Image Processing*, 28(3):1544–1555, 2019. 2

[5] D. H. Cho, Y. Matsushita, Y. W. Tai, and I. S. Kweon. Semi-calibrated photometric stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 2

[6] P. F. U. Gotardo, T. Simon, Y. Sheikh, and I. Matthews. Photogeometric scene flow for high-detail dynamic 3d reconstruction. In *IEEE International Conference on Computer Vision (ICCV)*, pages 846–854, 2015. 2

[7] C. Hane, C. Zach, A. Cohen, R. Angst, and M. Pollefeys. Joint 3D Scene Reconstruction and Class Segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013. 2

[8] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331, 1988. 2

[9] F. Lauze, Y. Quéau, and E. Plenge. Simultaneous reconstruction and segmentation of CT scans with shadowed data. In *International Conference on Scale Space and Variational Methods in Computer Vision (SSVM)*, pages 308–319, 2017. 2

[10] F. Logothetis, R. Mecca, and R. Cipolla. Semi-calibrated near field photometric stereo. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 941–950, 2017. 2

[11] F. Lu, X. Chen, I. Sato, and Y. Sato. SymPS: BRDF symmetry guided photometric stereo for shape and light source estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(1):221–234, 2017. 2

[12] R. Mecca, E. Rodolà, and D. Cremers. Realistic photometric stereo using partial differential irradiance equation ratios. *Computers & Graphics*, 51:8–16, 2015. 2

[13] Z. Mo, B. Shi, F. Lu, S.-K. Yeung, and Y. Matsushita. Uncalibrated photometric stereo under natural illumination. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2936–2945, 2018. 2

[14] D. Mumford and J. Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *Communications on pure and applied mathematics*, 42(5):577–685, 1989. 2

[15] C. Nieuwenhuis and D. Cremers. Spatially varying color distributions for interactive multi-label segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(5):1234–1247, 2013. 5

[16] S. Osher and J. A. Sethian. Fronts propagating with curvature-dependent speed: algorithms based on Hamilton-Jacobi formulations. *Journal of Computational Physics*, 79(1):12–49, 1988. 3

[17] W. A. Perkins. Area segmentation of images using edge points. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-2(1):8–15, 1980. 2

[18] V. A. Prisacariu, A. V. Segal, and I. Reid. Simultaneous monocular 2d segmentation, 3d pose recovery and 3d reconstruction. In *Asian Conference on Computer Vision*, pages 593–606, 2012. 2

[19] Y. Quéau, J.-D. Durou, and J.-F. Aujol. Normal integration: a survey. *Journal of Mathematical Imaging and Vision*, 60(4):576–593, 2018. 2

[20] Y. Quéau, R. Mecca, and J.-D. Durou. Unbiased photometric stereo for colored surfaces: A variational approach. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4359–4368, 2016. 1, 2

[21] Y. Quéau, T. Wu, F. Lauze, J.-D. Durou, and D. Cremers. A Non-Convex Variational Approach to Photometric Stereo under Inaccurate Lighting. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 350–359, 2017. 2

[22] S. Sengupta, H. Zhou, W. Forkel, R. Basri, T. Goldstein, and D. Jacobs. Solving uncalibrated photometric stereo using fewer images by jointly optimizing low-rank matrix completion and integrability. *Journal of Mathematical Imaging and Vision*, 60(4):563–575, 2018. 2

[23] B. Shi, Z. Mo, Z. Wu, D. Duan, S. 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. 1, 2, 4

[24] W. Smith and F. Fang. Height from photometric ratio with model-based light source selection. *Computer Vision and Image Understanding*, 145:128–138, 2016. 2

[25] K. Tateno, F. Tombari, and N. Navab. When 2.5 D is not enough: Simultaneous reconstruction, segmentation and recognition on dense SLAM. In *The IEEE International Conference on Robotics and Automation (ICRA)*, pages 2295–2302, 2016. 2

[26] R. J. Woodham. Photometric Method for Determining Surface Orientation from Multiple Images. *Optical Engineering*, 19(1):139–144, 1980. 1, 2

[27] L. Wu, A. Ganesh, B. Shi, Y. Matsushita, Y. Wang, and Y. Ma. Robust photometric stereo via low-rank matrix completion and recovery. In *Asian Conference on Computer Vision (ACCV)*, pages 703–717, 2010. 4

[28] Q. Zhang, R. Plemmons, D. Kittle, D. Brady, and S. Prasad. Joint segmentation and reconstruction of hyperspectral data with compressed measurements. *Applied Optics*, 50(22):4417–4435, 2011. 2

[29] Q. Zheng, B. Kumar, A. Shi, and G. Pan. Numerical Reflectance Compensation for Non-Lambertian Photometric Stereo. *IEEE Transactions on Image Processing*, 28(7):3177–3191, 2019. 2

# Chapter 8

# Recovering Real-World Reflectance Properties and Shading From HDR Imagery

## Copyright


Bjoern Haefner, Simon Green, Alan Oursland, Daniel Andersen, Michael Goesele, Daniel Cremers, Richard Newcombe, and Thomas Whelan
**Recovering Real-World Reflectance Properties and Shading From HDR Imagery**
*2021 International Conference on 3D Vision (3DV)*
DOI: 10.1109/3DV53792.2021.00115

## Individual contributions

Leading role in realizing the scientific project.

| | |
|---|---|
| Problem definition | *significantly contributed* |
| Literature survey | *significantly contributed* |
| Implementation | *significantly contributed* |
| Experimental evaluation | *significantly contributed* |
| Preparation of the manuscript | *significantly contributed* |

In accordance with the *IEEE Thesis / Dissertation Reuse Permissions*, we include the accepted version of the original publication [2] in the following.

# Recovering Real-World Reflectance Properties and Shading From HDR Imagery

Bjoern Haefner[1,2] Simon Green[2] Alan Oursland[2] Daniel Andersen[2]
Michael Goesele[2] Daniel Cremers[1] Richard Newcombe[2] Thomas Whelan[2]
[1]Technical University of Munich [2]Facebook Reality Labs Research
{bjoern.haefner, cremers}@tum.de,
{simongreen, ours, andersed, goesele, newcombe, twhelan}@fb.com

## Abstract

*We propose a method to estimate the bidirectional reflectance distribution function (BRDF) and shading of complete scenes under static illumination given the 3D scene geometry and a corresponding high dynamic range (HDR) video. By splitting the BRDF into its diffuse and non-diffuse parts we solve the estimation of each component separately. For the diffuse component, we sample the incident illumination at each point in the scene using Monte Carlo ray tracing, allowing us to factor the captured surface color into albedo and shading. We then use a novel ray tracing-based optimization strategy to estimate the non-diffuse parameters of the BRDF. In a variety of experiments, we demonstrate that our method efficiently generates realistic copies of the observed scenes.*

## 1. Introduction

Recovering a faithful copy of our world is of fundamental importance for virtual, augmented and mixed reality (VR, AR, MR). VR devices immerse the user into a virtual world to fulfill certain tasks, e.g. medical, educational or gaming purposes. They rely on a representation of the scene in terms of surface geometry, material properties and lighting. Since hand-crafting such virtual world models is tedious, there is an increasing demand for methods that can automatically reconstruct real world environments. Yet, their practical value critically depends on the realism of the virtual world. In MR and AR, faithful scene representations are required to render virtual objects that have the correct physical interactions and visual appearance with respect to their surroundings. While the reconstruction of surface geometry is quite mature, the estimation of reflectance and lighting of a scene remains a difficult open challenge – in particular, if we want to estimate these properties straight from an input video. This work brings the virtual and real world closer together enabling users to immerse in a *realistic virtual reality* and experience believable augmentations.

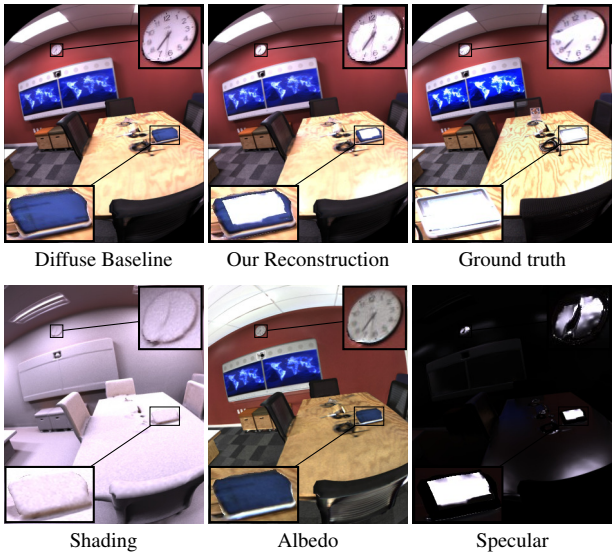



Figure 1. Reconstruction results: Given an input video and a geometric reconstruction of the scene, we deduce the scene's shading, albedo and specular properties, thereby allowing for a more faithful reconstruction. The insets show details of two specular objects.

Given a comprehensive HDR video of an environment and its corresponding reconstructed 3D mesh, we claim three novel contributions:

- An efficient method to leverage HDR textures for estimating albedo and shading per surface element.
- A procedure to calculate ideal target frames for each object in the scene within the estimation process.
- A method to estimate the non-diffuse BRDF using grid search with nested least-squares optimization.

On a broad range of real-world datasets, we demonstrate that this enables faithful reconstructions, plausible scene relighting and visually accurate rendering of virtual objects that can take the surrounding scene appearance and geometry into account.

## 2. Background and related work

In the following we recall the rendering equation [18] and discuss efforts to invert it in order to recover realistic models of the observed world.

### 2.1. The rendering equation

The rendering equation [18] is a useful and popular tool to render images given the scenes properties of material, illumination and geometry. It models the light transport as:

$$L_{\mathrm{o}}(\boldsymbol{x},\omega_{\mathrm{o}})=L_{\mathrm{e}}(\boldsymbol{x},\omega_{\mathrm{o}})+\int_{\mathcal{H}^2} f_{\mathrm{r}}(\boldsymbol{x},\omega,\omega_{\mathrm{o}})L(\boldsymbol{x},\omega)\langle\omega,\boldsymbol{n}\rangle \mathrm{d}\omega \quad (1)$$

where the $L_{\mathrm{o}}$ is the observed radiance at $\boldsymbol{x}\in\mathbb{R}^3$ in direction $\omega_{\mathrm{o}}\in\mathbb{S}^2$, with $\mathbb{S}^2$ being the 3D unit sphere. $L_{\mathrm{e}}(\boldsymbol{x},\omega_{\mathrm{o}})$ describes the amount of light emitted at $\boldsymbol{x}$ in direction $\omega_{\mathrm{o}}$ by a light source. The integral over the hemisphere $\mathcal{H}^2$ oriented by the surface normal $\boldsymbol{n}\in\mathbb{S}^2$ positioned at $\boldsymbol{x}$, integrates along all incident directions $\omega$. The integrand describes the interaction between material, light and geometry, where the BRDF $f_{\mathrm{r}}$ models the reflectance properties of a variety of materials. The radiance $L(\boldsymbol{x},\omega)$ describes the amount of incoming light at $\boldsymbol{x}$ from direction $\omega$. The geometric term $\langle\omega,\boldsymbol{n}\rangle$ models the spread of incident illumination over the surface at a given angle, where $\langle\cdot,\cdot\rangle:\mathbb{R}^3\times\mathbb{R}^3\rightarrow\mathbb{R}$ describes the dot product. Evaluating Eq. (1) can result in high quality renderings close to real-world images [18, 37], providing the relation between a captured image and its scene. To this end we identify for each pixel $\boldsymbol{p}\in\Omega\subset\mathbb{R}^2$ of the image $I:\Omega\rightarrow\mathbb{R}^3$, the conjugate 3D-point $\boldsymbol{x}$. And $\omega_{\mathrm{o}}$ is the normalized vector pointing from $\boldsymbol{x}$ to $\boldsymbol{p}$, $I(\boldsymbol{p})=L_{\mathrm{o}}(\boldsymbol{x},\omega_{\mathrm{o}})$.

### 2.2. Inverting the rendering equation

Inferring camera and scene properties by inverting the rendering equation in order to obtain suitable models of the real world has a long-standing history and is called inverse rendering [35]. We now discuss the most related work that tackles the challenging task of material estimation, but refer to [20] for a comprehensive survey on inverse rendering.

**Deep Learning** [4, 5, 9, 24, 25, 26, 30, 43, 45, 54] approaches train a network in an (un-)supervised manner and demonstrate impressive results in the context of photorealistic scene reconstruction. Yet, these techniques applied in the single image domain, are concerned with single object reconstruction, and/or have an implicit scene representation. This makes it difficult to be compatible with conventional computer graphics assets used for lighting and physics interactions in full 3D room-scale real-world reconstructions, thus limiting the applicability of these approaches to AR/VR/MR applications.

**Multi-Shot** [1, 3, 4, 9, 11, 13, 21, 22, 26, 23, 27, 29, 30, 31, 33, 40, 42, 52, 53] techniques recover material effects using multiple images taken from the same or different viewpoints. While more observations constrain the resulting optimization problem better, additional images need to be captured and the computational burden can limit inference in terms of memory and runtime. Thus, it is always desirable to use as few images as possible, while still constraining the search space of possible solutions enough to find reasonable estimates. Additionally, many of these approaches have at best a piece-wise constant material per object if no active lighting is used.

**Active Lighting** [1, 9, 11, 13, 16, 15, 21, 22, 31, 36, 38, 42, 53] frameworks estimate reflectance properties similar to multi-shot techniques, but additionally require different (calibrated) illumination for each image. This limits the practicability of these approaches as a light source has to be actively controlled. It is known that these approaches are well-posed in the Lambertian setting under general illumination [6] and a point-wise solution of the albedo can be found as it is much more constrained. Considering view-dependent material effects adds additional complexity to the problem and even if illumination and geometry is known, recovering non-diffuse reflectance is an open challenge and additional assumptions have to be made [14].

**HDR Imagery** [1, 11, 21, 22, 29, 52, 53] shows great usage in photometric approaches as they tend to relate scene properties to linear radiance data instead of non-linearly mapped pixel intensities [12] – a relation which, if violated due to no camera calibration, can result in undesired deterioration [12, 19]. Interestingly, despite its potential and desirable properties the literature applying HDR data in the context of photorealistic reconstruction of room-sized environments is fairly sparse [29, 52]. This might be due to the different orders of magnitude involved when using HDR data – an effect non-existent with 8-bit images as higher radiance values are usually clamped to 255. This can cause standard algorithms, like running average of pixel intensities to not work as expected.

In contrast to the above approaches, the method presented here does not rely on large amounts of diverse training data, nor on active lighting and works in complete room-sized 3D environments. We effectively incorporate the advantages of HDR imagery and a tailored ray tracing framework to recover the BRDF for every object in the scene. More specifically, we can recover a spatially varying albedo, and present a principled way to leverage HDR video for automated target frame selection which allows us to estimate non-diffuse material effects from a single view per object. To the best of our knowledge this is the first work utilizing HDR data with consistent full 3D room-scale reconstructions, which is able to recover BRDF parameters of every object in the scene using a single automatically computed target frame. In the context of AR/VR/MR, this enables the faithful recovery of large-scale scenes that support conventional physical as well as light interaction between real and virtual objects.

## 3. Recovering complex reflectance and shading

Given a mesh-based 3D reconstruction of the scene geometry, an HDR RGB sequence of frames covering that geometry and their corresponding poses, we first reconstruct and estimate the lit diffuse HDR texture (Section 3.2). This then builds the foundation for the albedo and shading estimation using only the textured geometry (Section 3.3), and, given an object segmentation, the estimation of the specular material parameters per object (Section 3.4). See Algorithm 1 for an overview of our proposed framework. Note that our input assumptions differ only in the HDR data compared to other approaches like [3], allowing us to cover the dynamic range of the scene from the darkest to the brightest areas. We follow [12] to transform the captured data to floating point linear units directly proportional to the incoming radiance and discuss in Section 3.2 and 3.4.1 arising issues and how to effectively leverage that to our advantage.

**Algorithm 1** Overview of our proposed algorithm

**Input:** HDR data, poses, geometry, object segmentation
**Output:** $\tilde{\boldsymbol{\rho}}$, $\{\varphi^i, \psi^i\}_{i=1,\dots,M}$ for all $M$ objects in the scene

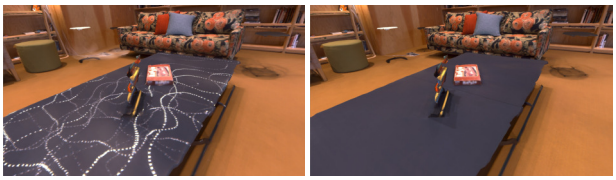
*Calculate lit diffuse HDR texture* (Sec. 3.2):
1: $L_\mathrm{d}$ = runningMedian(HDR data, poses, geometry)
*Calculate shading $S$ and albedo $\tilde{\boldsymbol{\rho}}$* (Sec. 3.3):
2: $S$ = calcShading(geometry, $L_\mathrm{d}$)
3: $\tilde{\boldsymbol{\rho}} = \frac{L_\mathrm{d}}{S}$
*For each object: Target frame calculation and roughness $\varphi^i$ and specular $\psi^i$ estimation* (Sec. 3.4):
4: TFs = calcTargetFrames(HDR data, poses, geometry, object segmentation)
5: **for** each object $i$ in the scene **do**
6: TF = TFs[i] ($i$-th object's target frame)
7: $\varphi^i$, $\psi^i$ = estimateNondiffuse(TF, geometry, $L_\mathrm{d}$)

### 3.1. Microfacet BRDF Model

We restrict our focus to isotropic, dielectric (non-metallic), and opaque (not translucent/transparent) objects only. A desirable property for a BRDF is an additive separation into its diffuse and non-diffuse component, as it allows splitting the problem of BRDF parameter estimation into two separate, easier to solve problems as we will describe later. We will thus use a dichromatic BRDF [44],

$$f_\mathrm{r}(\boldsymbol{x}, \omega, \omega_\mathrm{o}) = f_\mathrm{d}(\boldsymbol{x}) + f_\mathrm{nd}(\boldsymbol{x}, \omega, \omega_\mathrm{o}) \tag{2}$$

and identify the diffuse part as $f_\mathrm{d}(\boldsymbol{x}) = \frac{\boldsymbol{\rho}(\boldsymbol{x})}{\pi} =: \tilde{\boldsymbol{\rho}}(\boldsymbol{x})$, and call $\tilde{\boldsymbol{\rho}}$ the *(scaled) albedo*, where $\boldsymbol{\rho} : \Sigma \to [0,1]^3$, given a reconstructed surface $\Sigma \subset \mathbb{R}^3$. The non-diffuse component is described using the Torrance-Sparrow microfacet model [10, 49] with a GGX distribution [46, 50, 51] and



Running Mean Running Approximated Median

Figure 2. The mean textures (left) suffer from occluding edge bleeding and baked in specularities, our approximated median (right) is able to estimate textures without such artifacts.

Schlick's Fresnel approximation [41] (dropping the $\boldsymbol{x}, \omega, \omega_\mathrm{o}$ dependencies for brevity),

$$f_\mathrm{nd}(\varphi, \psi) = G(\varphi)\, D(\varphi)\, F(\psi), \tag{3}$$

$$G(\varphi) = G_1(\langle \boldsymbol{n}, \omega \rangle, \tilde{\varphi}) \cdot G_1(\langle \boldsymbol{n}, \omega_\mathrm{o} \rangle, \tilde{\varphi}), \tag{4}$$

$$D(\varphi) = \frac{\hat{\varphi}^2}{\pi \left(1 + (\hat{\varphi}^2 - 1)\langle \boldsymbol{n}, h \rangle^2\right)^2}, \tag{5}$$

$$F(\psi) = \tilde{\psi} + \left(1 - \tilde{\psi}\right)(1 - \langle \omega, h \rangle)^5, \tag{6}$$

with $G_1(x, y) = (x + \sqrt{x^2 + y^2 - x^2 y^2})^{-1}$, the half vector $h = \frac{\omega + \omega_\mathrm{o}}{\|\omega + \omega_\mathrm{o}\|}$, and the nondiffuse parameters *roughness* $\varphi : \Sigma \to [0,1]$, and *specular* $\psi : \Sigma \to [0,1]$. Following [7], we apply three reparameterisations to increase robustness: $\tilde{\varphi} = (\frac{\varphi}{2} + \frac{1}{2})^2$ to have a more perceptually linear change in roughness, $\hat{\varphi} = \max(0.001, \varphi)$ for numerical stability, and $\tilde{\psi} = 0.08\psi$ causing the refractive index to cover most common materials. Plugging (2) into (1), assuming non-emissivity ($L_\mathrm{e} \equiv 0$) and splitting the integral, we get

$$L_\mathrm{o}(\boldsymbol{x}, \omega_\mathrm{o}) = L_\mathrm{d}(\boldsymbol{x}) + L_\mathrm{nd}(\boldsymbol{x}, \omega_\mathrm{o}), \tag{7}$$

$$L_\mathrm{d}(\boldsymbol{x}) := f_\mathrm{d}(\boldsymbol{x}; \boldsymbol{\rho}) \int_{\mathcal{H}^2} L(\boldsymbol{x}, \omega) \langle \omega, \boldsymbol{n} \rangle \, \mathrm{d}\omega, \tag{8}$$

$$L_\mathrm{nd}(\boldsymbol{x}, \omega_\mathrm{o}) := \int_{\mathcal{H}^2} f_\mathrm{nd}(\boldsymbol{x}, \omega, \omega_\mathrm{o}; \varphi, \psi) L(\boldsymbol{x}, \omega) \langle \omega, \boldsymbol{n} \rangle \, \mathrm{d}\omega. \tag{9}$$

### 3.2. Lit diffuse HDR texture estimation

We estimate the *lit diffuse HDR texture* by projecting the video frames onto the surface geometry. Using low dynamic range 8-bit data, weighted averaging [8, 32] typically yields reasonable results as outliers are smoothed out. However, this is not the case with HDR data due to its large range of values, resulting in a number of visual artifacts caused by two main phenomena: errors in the geometry and bright lights along with specular reflections of those, see Fig. 2 left. One popular approach to diminish these artifacts is to calculate the median rather than a running mean [39]. However, this is extremely expensive since it requires storing all RGB values. To overcome this we estimate an approximation of the median of each color channel using the P-Square

algorithm [17][1]. Figure 2 shows a comparison between the running mean and our running approximated median. Despite errors in the reconstruction, the floor is no longer corrupted and specular reflections on the table have been removed. Mathematically speaking, the BRDF inscribed in the texture should have no view-dependent effects and can thus be assumed to represent the albedo. Nevertheless, one should not identify the median texture with the albedo itself as it still contains global light transport and geometric information. Thus, we assume that the median texture can be identified as the diffuse radiance, $L_\mathrm{d}$ and we call it the *lit diffuse HDR texture*.

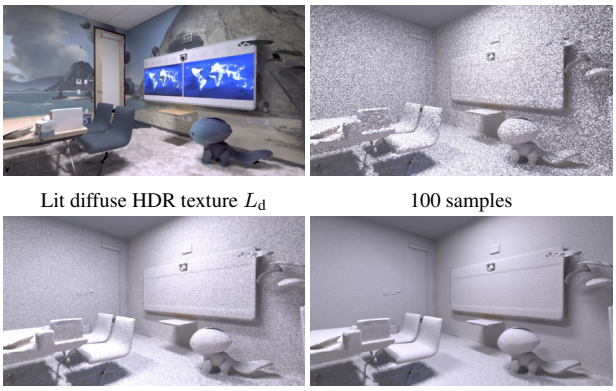
### 3.3. Albedo and shading estimation

We are now going to effectively leverage the information that the HDR texture's intensity is proportional to the true radiance, which would not be possible with textures where intensities, especially at light sources, are truncated to 8-bits. This allows us to estimate the captured shading $S$ at each surface point $\boldsymbol{x} \in \Sigma$ of the scene,

$$S(\boldsymbol{x}) := \int_{\mathcal{H}^2} L(\boldsymbol{x}, \omega) \langle \omega, \boldsymbol{n} \rangle \, \mathrm{d}\omega. \tag{10}$$

The shading describes the sum of the radiance $L(\boldsymbol{x}, \omega)$ gathered from the scene weighted by the geometric scale factor $\langle \omega, \boldsymbol{n} \rangle$. We estimate the shading $S$ via Monte-Carlo ray tracing, a stochastic approach to estimate complex integrals such as Eq. (10). We cast rays at each scene's surface point $\boldsymbol{x} \in \Sigma$ on the hemisphere $\mathcal{H}^2$, where the chosen ray directions $\omega$ follow a distribution accounting for the scalar product in Eq. (10) (cosine weighted) [37]. For each cast ray $(\boldsymbol{x}, \omega)$ we read the lit diffuse HDR texture at the closest hit point $\tilde{\boldsymbol{x}}$ and interpret it as the incident radiance, $L(\boldsymbol{x}, \omega) = L_\mathrm{d}(\tilde{\boldsymbol{x}})$. Summing up all cosine weighted samples of incident radiance gives an estimate for the shading $S$ for each surface point $\boldsymbol{x}$. One can interpret this procedure as sampling each surface point's environment map. Note that the captured lit diffuse HDR texture already includes the effects of global light transport in the diffuse scene [52], thus we can perform the proposed shading estimation in parallel for all surface points $\boldsymbol{x} \in \Sigma$ independently. Finally, as the captured lit diffuse HDR texture is the product of the scaled albedo and the shading, see Eq. (8), we can recover the albedo by dividing the captured lit diffuse HDR texture by the estimated shading.

Fig. 3 shows shading estimates for different numbers of ray samples and how increasing sample size de-noises the result. Our approach to recover shading and albedo does not account for emissive radiance $L_\mathrm{e}$, although the lit diffuse HDR textures inherently carries that information. Nevertheless, we do not think of this as a major disadvantage, as for emissive objects, the impact of the intrinsic radiance (what we see when looking at a light source which is turned off) is negligible compared to its emissive radiance.



Lit diffuse HDR texture $L_\mathrm{d}$ | 100 samples

1000 samples | 10000 samples

Figure 3. Estimated shading $S$ for different numbers of ray samples. Note how more samples remove noise from the shading.

### 3.4. Specular appearance estimation

Given an estimate of the lit diffuse scene $L_\mathrm{d}(\boldsymbol{x})$ at each surface point, we can estimate the non-diffuse BRDF parameters $\psi$ and $\varphi$ per object. We assume that for all $M$ objects in the scene, each object's view-dependent effects can be described with two parameters, $\{(\psi^i, \varphi^i)\}_{i=1,\dots,M}$, resulting in two constant, non-diffuse BRDF parameters per object. We first discuss how we automatically select an individual target frame per object given an object segmentation before utilizing these in the proposed optimization scheme.

#### 3.4.1 Target frames

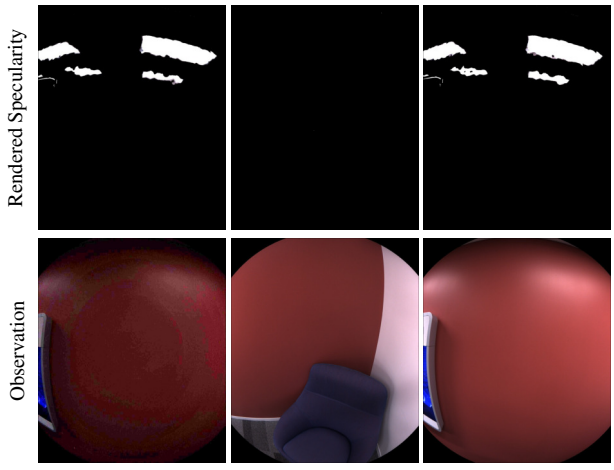
Path tracing a single image is expensive and time-consuming, which is why we would like to use as few images for inference as possible. Additionally, the so called target frames (TF) used to estimate each object's non-diffuse material parameters should have two attributes:

- $\mathcal{A}_1$, high chance of specular highlights caused by direct illumination, and
- $\mathcal{A}_2$ the captured observation from the HDR video consists mostly of valid pixels, i.e., the RGB values are not over- or under-saturated[2].

These requirements in combination with the assumption that a single object's specular appearance can be described with two parameters allows us to use *only one* TF per object. In order to find TFs fulfilling $\mathcal{A}_1$, we assume the object of interest is a perfect mirror and we render only the pixels where light sources can be seen in the mirrored surface.

[1] In the interest of brevity we refer to the original paper for a full description of the algorithm and its performance relative to an exact median.

[2] Over- or under-saturated observations do not depend linearly on the incoming radiance [12], hence we omit them to avoid corrupting the result.




| | | | |
|---|---|---|---|
| $\mathcal{A}_1$ | ✓ | ✗ (no highlights) | ✓ |
| $\mathcal{A}_2$ | ✗ (under-saturated) | ✓ | ✓ |
| TF | ✗ | ✗ | ✓ |

Figure 4. Example of good and bad target frame (TF) candidates for the object "Red Wall" based on the attributes $\mathcal{A}_1$ and $\mathcal{A}_2$. While for the first two columns, either the observation is under-saturated (intensity increased by a factor of 10 for visualization purposes) or there are no specular highlights on the object, the third column shows a TF satisfying both $\mathcal{A}_1$ and $\mathcal{A}_2$.

Note that this is the only step in our framework that requires information about position of emitting light sources. Concerning $\mathcal{A}_2$, the HDR capture cycles through three different exposures in subsequent frames. This leads to three different exposures at roughly the same viewpoint, allowing us to find at least one frame with enough valid pixels. Example frames and their attributes $\mathcal{A}_1$ and $\mathcal{A}_2$ are visualized in Fig. 4. We iterate through the video and for each of it's objects set the TF as the frame with most pixels in $\mathcal{A}_1 \cap \mathcal{A}_2$.

### 3.4.2 Optimization

Given the $i$-th object's target frame $I^i$ we describe it as the composition of its diffuse and non-diffuse component, $I^i_{\text{d}}$ and $I^i_{\text{nd}}$ respectively,

$$I^i\left(\boldsymbol{p}\right) = I^i_{\text{d}}\left(\boldsymbol{p}\right) + I^i_{\text{nd}}\left(\boldsymbol{p}; \varphi^i, \psi^i\right). \tag{11}$$

We assume $I^i$ (the observation) and $I^i_{\text{d}}$ (rendered image using the lit diffuse HDR texture) to be given so that the only varying quantity is $I^i_{\text{nd}}$. Due to view-dependent appearance effects, full evaluation, i.e., dense sampling of $\omega$ of the integral in Eq. (9) is challenging. We therefore follow a multi importance sampling strategy [37]. For further technical details see the supplementary material. We assume that single bounce ray tracing is enough for a reasonable approximation of the scene [52] keeping computational expense practical. We have a good estimation of the lit diffuse scene thanks to the HDR median textures $L_{\text{d}}$. When adding view-dependent effects such as reflections, inter-reflections start to have impact on the final result. Nevertheless, a specular lobe illuminating the scene is assumed to be negligible compared to an emissive light source when integrating over the whole hemisphere, as our target frames are chosen such that specular reflections are mainly caused by direct illumination ($\mathcal{A}_1$). Even in the presence of mirror like objects our target frames were not corrupted enough with indirect illumination that this would cause the system to fail. In order to determine the non-diffuse properties of the BRDF we can now formulate an optimization problem in $\mathcal{X}^i := (\varphi^i, \psi^i)$ per object $i$, i.e., we want to solve for $i = 1, \ldots, M$,

$$\min_{\mathcal{X}^i \in [0,1]^2} \mathcal{L}(\mathcal{X}^i) := \sum_{\boldsymbol{p} \in \Omega^i} \left\| r\left(\boldsymbol{p}; \mathcal{X}^i\right) \right\|_2^2. \tag{12}$$

$\|\cdot\|_2$ is the $L_2$-norm and $r$ a point-wise RGB-color residual at each pixel $\boldsymbol{p}$ in the image domain $\Omega^i \subset \Omega$ showing only the $i$-th object,

$$r\left(\boldsymbol{p}; \mathcal{X}^i\right) = I^i\left(\boldsymbol{p}\right) - \left(I^i_{\text{d}}\left(\boldsymbol{p}\right) + \mathcal{I}^i_{\text{nd}}\left(\boldsymbol{p}; \mathcal{X}^i\right)\right). \tag{13}$$

Note that due to the single bounce assumption, the $M$ optimization problems in (12) are disjoint, which enables solving each problem separately and in parallel.

Optimization problems like Eq. (12) are difficult to solve due to the non-convexity in the roughness parameter $\varphi^i$, cp. Eqs. (4) and (5). We now present a simple and fast numerical scheme that can tackle the inherent complexity by exploiting the closed parameter domain $[0, 1]^2$ of $\mathcal{X}^i$ and the fact that the BRDF $f_{\text{nd}}$ depends only linearly on the specular parameter $\psi^i$, cp. Eq. (6). We perform a two-level grid search approach (in $\varphi^i$) with nested least squares optimization (in $\psi^i$). That is, at the $l$-th level we set the roughness $\varphi^i = \varphi^i_{l_k}$ from a discrete set of sample points equidistantly spread across an interval $[a_l, b_l]$,

$$\varphi^i_{l_k} \in \{a_l + \frac{k \cdot (b_l - a_l)}{K - 1} \mid k = 0, \ldots, K-1\} \tag{14}$$

For each $\varphi^i_{l_k}$ we calculate the best specular value $\psi^i_{l_k}$ by solving the linear least squares problem in Eq. (12) in closed form. For the resulting $K$ tuples $\{\mathcal{X}^i_{l_k}\}_{k=0,\ldots,K-1}$ at the $l$-th level we evaluate $\mathcal{L}(\mathcal{X}^i_{l_k})$ and set the minimizer of the $l$-th level as the one with lowest cost. We choose $K = 11$, as we found this gives a dense enough sampling of the interval $[a_l, b_l]$ $\forall l$. At the first level we set $a_0 = 0$, $b_0 = 1$, while the second level's interval is initialized with the direct left and right neighbours of the minimizer's roughness value, or the roughness value itself in case it lies on the boundary of the sampling interval. Note that this approach always terminates after $K \cdot$"number of levels"$= 22$ iterations, but is not guaranteed to find a global minimizer – a challenging task in non-convex optimization. In our evaluation we did not observe any failed results that were undoubtedly caused by an unsuccessful optimization.

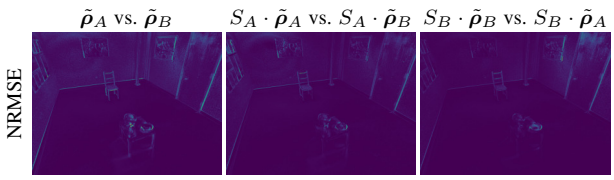



Figure 5. Error maps and numbers of normalized RMSE (NRMSE) of our approach to estimate albedo and shading on two differently illuminated scans $A$ and $B$. We compare the two estimated albedos $\tilde{\rho}_A$ and $\tilde{\rho}_B$, and how well the ground truth ($S_A \cdot \tilde{\rho}_A$ and $S_B \cdot \tilde{\rho}_B$) can be predicted with the other scan's albedo ($S_A \cdot \tilde{\rho}_B$ and $S_B \cdot \tilde{\rho}_A$). See Figure 6 for the images used to calculate the shown error maps.

## 4. Experiments

Given each surface point's albedo and shading, as well as each object's specular appearance, we can now quantitatively and qualitatively evaluate the effectiveness of our proposed approach. We use the Replica dataset [48] for the evaluation as this provides appropriate input data: reconstructed meshes of the complete scene, HDR video (provided by the authors of [48]), per frame camera poses, and semantic object instance information.

For quantitative validation of the albedo and shading estimation we use a dataset captured by ourselves with control over illumination and the objects in the scene, see the supplementary material for details on the capturing process. The room has in total four globe lights and three LED panels as light sources, which differ in wavelength and emission. The two scans differ in their respective lighting: For the first scan, all four globe lights and one LED panel were turned on (we call this *Scan/Reconstruction* $A$). For the second scan, only two LED panels (different from the one in *Scan* $A$) were turned on (we call this *Scan/Reconstruction* $B$). Note that we calculate the set of lit diffuse HDR textures (Section 3.2) a priori for each dataset, which runs on the GPU at $\approx$8–9Hz for RGB images of resolution $1224 \times 1024$. All experiments are carried out on a machine with an Intel Xeon 3.70GHz and an NVIDIA GeForce RTX 2080. *We encourage the reader to view our supplementary material for further results.*

### 4.1. Albedo and shading validation

We use *Reconstructions* $A$ and $B$ as well as scenes from the Replica dataset [48] to evaluate the albedo and shading described in Section 3.3. Using NVidia's OptiX engine [34], we cast 10000 rays per texel from each corresponding surface element to get a de-noised estimate of the shading and albedo. This process takes $\sim$10min.

**Quantitative evaluation** is carried out on the *Reconstructions* $A$ and $B$. The reconstructed albedos should ideally be equal as lighting cues are explained by the shading $S$. Figure 5 shows the normalized RMSE (NRMSE) for the whole scene verifying that there is only little, i.e. less than $10\%$ difference between the two albedos. Additionally, a numerical evaluation between the ground truth and predicted reconstructions is carried out. To this end, we compare $\tilde{\rho}_A \cdot S_A$ vs. $\tilde{\rho}_B \cdot S_A$ to see how well reconstruction $A$ can be predicted, while $\tilde{\rho}_B \cdot S_B$ vs. $\tilde{\rho}_A \cdot S_B$ validates the prediction of reconstruction $B$. An error well below $5\%$ for both tests shows that we can faithfully modify diffuse scenes with novel lighting conditions. Figure 6 shows the estimated albedos, shadings, ground truth and their predictions. While overall both albedo estimates are visually almost identical, few artifacts are visible and show how our system performs under violated assumptions of 1) remaining view-dependent effects in the lit diffuse HDR texture (e.g. on the door), and 2) inaccuracies in the reconstructed geometry (e.g. the objects on the table). Nevertheless, as numerically and quali-

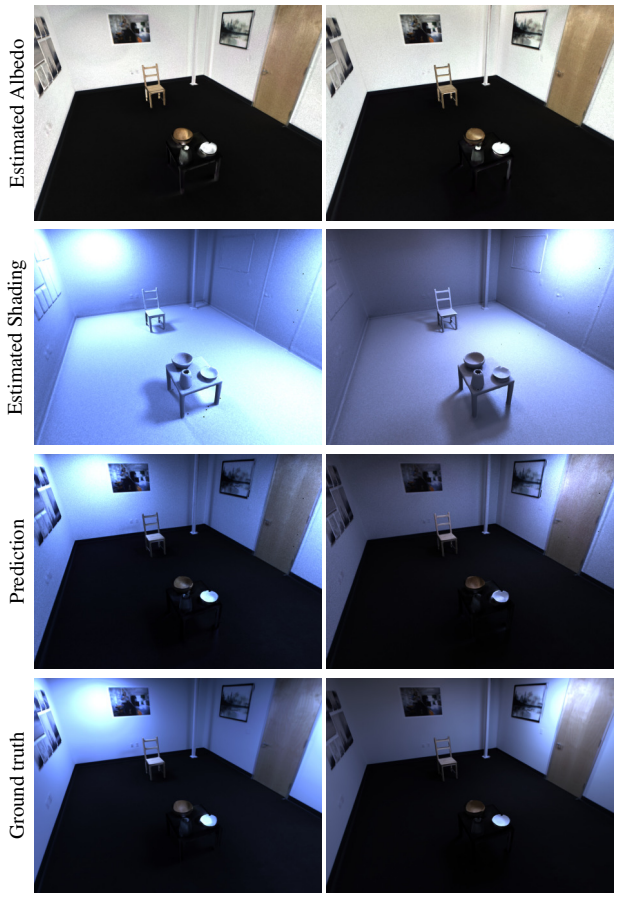
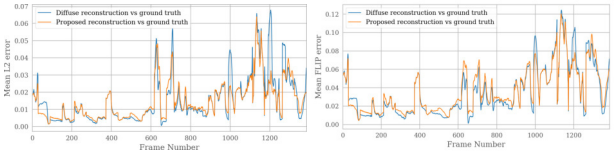



Figure 6. Numerical evaluation of our approach to estimate albedo and shading on two differently illuminated scans $A$ and $B$. As can be seen visually, illumination information is nicely explained in the estimated shading, while both albedos look almost identical and the predicted scene is close to ground truth.


Input Albedo estimate Shading estimate

Figure 7. We deploy our albedo and shading estimation on challenging real-world "Office" data sets of the Replica data set [48] and are able to estimate per-texel albedo and shading information, using the reconstructed mesh and lit diffuse HDR texture only. More results can be found the in the supplementary material.




Mean $L_2$ error Mean FLIP error

Figure 8. Numerical comparison on Office 1 [48] between a purely diffuse reconstruction with the ground truth (blue line) and the proposed reconstruction with the ground truth (orange). The left shows the numerical mean $L_2$ metric, while the right visualises the perceptual FLIP [2] metric. More results can be found in the supplementary material.

tatively shown, errors in our albedo estimation are still tolerable to plausibly relight diffuse scenes, i.e. errors between the two albedo estimates are easier to detect than errors between predicted relighting and ground truth.

**Qualitative evaluation** is carried out on the real-world Replica dataset [48] and can be seen in Figure 7. When our assumptions are met, we can recover an albedo estimate free of illumination effects, as these are contained in the corresponding shading estimate. Furthermore, we are able to tackle the challenging task of removing cast shadows of objects, e.g., chairs, sofas and tables. Note that in Office 0 the table under the display has a stand right below it on the floor which can be mistaken as a cast shadow in the albedo estimate, but the corresponding shading estimate shows it has actually been successfully removed.

### 4.2. Specular appearance estimation validation

For quantitative and qualitative comparison, we deploy our approach described in Section 3.4 on the Replica dataset [48]; casting 200 rays per each pixel's corresponding surface element using OptiX [34]. The dataset consists of $\approx$50–150 objects per scene; each of different size, geometry and material. Estimating an object's non-diffuse BRDF parameters takes $\approx$238sec on the GPU.

**Quantitative evaluation** is concerned with how much a reconstruction improves compared to the *diffuse baseline*, i.e. a reconstruction using the lit diffuse HDR textures. We infer non-diffuse material parameters from a single image per object. More specifically, to validate consistency across different views we test our predictions against unseen viewpoints of the ground truth observation, and compare this to the diffuse baseline. To this end we use two different error metrics, the numerical $L_2$-loss, as well as the recently introduced perceptual FLIP [2] evaluator. FLIP has a particular focus on the differences between rendered images and corresponding ground truths via approximating the difference perceived by humans when alternating between two images. Figure 8 shows the per frame numerical mean $L_2$ metric (left), and the perceptual mean FLIP [2] metric (right) for a video sequence of Office 1 [48] containing 1389 frames, where 1363 frames are novel viewpoints and only 26 frames were used as target frames. Both graphs show that on average the error decreases when incorporating the proposed view-dependent BRDF estimates. Note that, besides only small differences between the orange and blue graph (as specular highlights are only sparsely distributed across an image, if they appear at all), the improvements ("orange<blue") are of much larger magnitude than the deterioration ("blue<orange"). That means that if our proposed rendering degrades the ground truth more than the diffuse baseline, it is only slightly worse, while our proposed rendering considerably improves realism compared to the diffuse baseline.

**Qualitative evaluation and comparison to related work** is carried out over multiple real-world datasets of [48], see Figure 9. The state-of-the-art approach closest related to ours is [3], which is a full path tracing (2 bounces) approach to estimate the scene's material properties. We chose the hyper-parameters as recommended by the authors using 1 sample to estimate the image and 511 for the derivative and ran [3] until convergence which took 12−24hrs, depending on the data set. A side-by-side comparison between the diffuse, the state-of-the-art [3] and the proposed reconstruction along with the ground truth and the corresponding error maps shows the superiority of our approach. While the overall trend of estimated material parameters of [3] seems correct, Monte Carlo noise is dominating the resulting reconstruction which heavily deteriorates the rendered images. Our method can successfully reconstruct subtle specular effects such as the specular lobe on the wall in Office 0. It also models stronger reflections, e.g., the TV screen in Office 4 and even mirror like reflections, see the glass window in Office 3. Inaccuracies in geometry can affect the result (Office 1, largest deterioration according to Fig. 8), although the tablet glass screen seems to be well estimated.

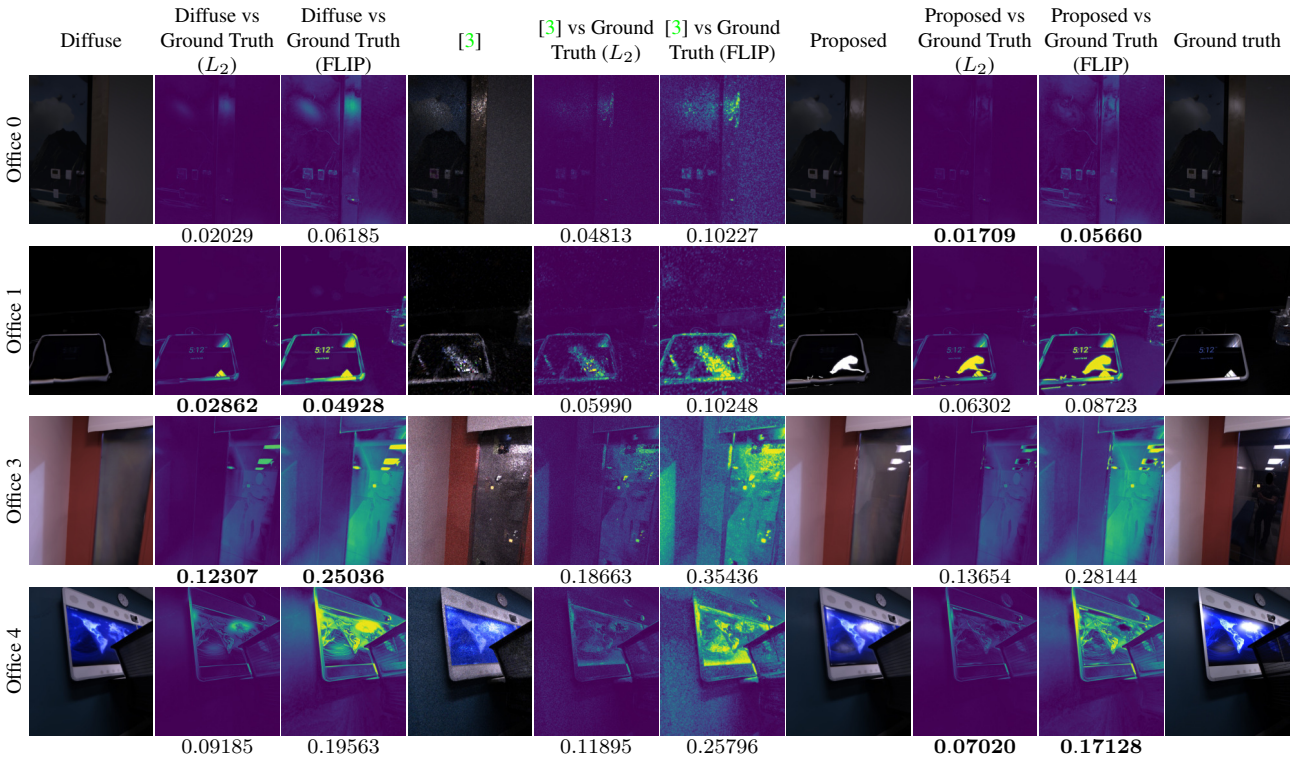



Figure 9. Side-by-side comparisons between the diffuse baseline, a path tracing approach [3] and the proposed reconstruction along with the ground truth and the corresponding $L_2$ errors and FLIP evaluator [2]. More results can be found in the supplementary material.

Figure 10. Complete synthetic relighting of different data sets (Office 0, Office 2 of [48]) with virtually placed objects [47, 28]. More results can be found in the supplementary material.

### 4.3. Relighting

Finally, having the full BRDF at hand (albedo, specular, and roughness), we can now do a complete visually accurate rendering of the full scene under new synthetic lighting with additional virtual objects, see Figure 10. To this end, we deploy a path tracing engine with four bounces. The bunny and statue added to the reconstructions of the Office 0 and Office 2 scenes [48] look faithful and realistic as they take the overall scene's appearance into account resulting in consistent shadowing and material effects.

### 4.4. Limitations and future work

We assume geometry to be given, thus deterioration can have negative impact on the result (Figure 9 Office 1) – a standard limitation for inverse rendering under known geometry [3, 13, 52]. In our tests, we did not experience inter-reflections to cause our system to fail, as target frames are chosen to maximize specular reflections based on direct illumination. However, we expect the presence of strong inter-reflections to limit the performance of our framework, due to the single bounce assumption. In the future, we aim to overcome some limitations with the lit diffuse HDR texture, as it can suffer from remaining baked-in view-dependent effects. While modest corruptions are tolerable and still enable plausible relighting (Section 4.1), we assume the system to not work as assumed when artifacts dominate the median texture.

## 5. Conclusion

We introduced a method that estimates the BRDF and shading properties of complete 3D scenes from HDR imagery. We are able to recover per surface element albedo and shading using only the reconstructed geometry and HDR textures. We provide a scheme to automatically calculate target frames per object; these are then used to estimate non-diffuse material parameters per object. Numerous experiments on a range of challenging real-world HDR data sets validate the efficiency of our approach compared to the current state-of-the-art, allowing us to create reconstructions that are almost indistinguishable from the real-world.

## References


[1] N. Alldrin, T. Zickler, and D. Kriegman. Photometric stereo with non-parametric and spatially-varying reflectance. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 2

[2] P. Andersson, J. Nilsson, T. Akenine-Möller, M. Oskarsson, K. Åström, and M. D. Fairchild. Flip: a difference evaluator for alternating images. *Proceedings of the ACM on Computer Graphics and Interactive Techniques (HPG 2020)*, 3(2), 2020. 7, 8

[3] D. Azinovic, T.-M. Li, A. Kaplanyan, and M. Niessner. Inverse path tracing for joint material and lighting estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2, 3, 7, 8

[4] S. Bi, Z. Xu, P. Srinivasan, B. Mildenhall, K. Sunkavalli, M. Hašan, Y. Hold-Geoffroy, D. Kriegman, and R. Ramamoorthi. Neural reflectance fields for appearance acquisition. *arXiv preprint arXiv:2008.03824*, 2020. 2

[5] S. Bi, Z. Xu, K. Sunkavalli, D. Kriegman, and R. Ramamoorthi. Deep 3d capture: Geometry and reflectance from sparse multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5960–5969, 2020. 2

[6] M. Brahimi, Y. Quéau, B. Haefner, and D. Cremers. *On the Well-Posedness of Uncalibrated Photometric Stereo Under General Lighting*, chapter Advances in Photometric 3D-Reconstruction, pages 147–176. Springer International Publishing, Cham, 2020. 2

[7] B. Burley and W. D. A. Studios. Physically-based shading at disney. In *ACM SIGGRAPH*, volume 2012, pages 1–7, 2012. 3

[8] E. Bylow, J. Sturm, C. Kerl, F. Kahl, and D. Cremers. Real-time camera tracking and 3d reconstruction using signed distance functions. In *Robotics: Science and Systems Conference (RSS)*, June 2013. 3

[9] C. Che, F. Luan, S. Zhao, K. Bala, and I. Gkioulekas. Inverse transport networks. *arXiv preprint arXiv:1809.10820*, 2018. 2

[10] R. L. Cook and K. E. Torrance. A reflectance model for computer graphics. *ACM Transactions on Graphics (ToG)*, 1(1):7–24, 1982. 3

[11] P. Debevec. Rendering synthetic objects into real scenes: Bridging traditional and image-based graphics with global illumination and high dynamic range photography. In *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '98, page 189–198, New York, NY, USA, 1998. Association for Computing Machinery. 2

[12] P. E. Debevec and J. Malik. Recovering high dynamic range radiance maps from photographs. In *Proceedings of SIGGRAPH*, 1997. 2, 3, 4

[13] Y. Dong, G. Chen, P. Peers, J. Zhang, and X. Tong. Appearance-from-motion: Recovering spatially varying surface reflectance under unknown lighting. *ACM Transactions on Graphics (TOG)*, 33(6):1–12, 2014. 2, 8

[14] D. Gao, X. Li, Y. Dong, P. Peers, K. Xu, and X. Tong. Deep inverse rendering for high-resolution svbrdf estimation from an arbitrary number of images. *ACM Transactions on Graphics (TOG)*, 38(4):1–15, 2019. 2

[15] B. Haefner, S. Peng, A. Verma, Y. Quéau, and D. Cremers. Photometric depth super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(10):2453–2464, 2020. 2

[16] B. Haefner, Z. Ye, M. Gao, T. Wu, Y. Quéau, and D. Cremers. Variational uncalibrated photometric stereo under general lighting. In *International Conference on Computer Vision (ICCV)*, Seoul, South Korea, October 2019. 2

[17] R. Jain and I. Chlamtac. The $p^2$ algorithm for dynamic calculation of quantiles and histograms without storing observations. *Commun. ACM*, 1985. 4

[18] J. T. Kajiya. The rendering equation. In *Proceedings of the 13th annual conference on Computer graphics and interactive techniques*, pages 143–150, 1986. 2

[19] H. C. Karaimer and M. S. Brown. A software platform for manipulating the camera imaging pipeline. In *European Conference on Computer Vision (ECCV)*, 2016. 2

[20] H. Kato, D. Beker, M. Morariu, T. Ando, T. Matsuoka, W. Kehl, and A. Gaidon. Differentiable rendering: A survey. 2020. 2

[21] H. P. Lensch, J. Kautz, M. Goesele, W. Heidrich, and H.-P. Seidel. Image-based reconstruction of spatially varying materials. In *Eurographics Workshop on Rendering Techniques*, pages 103–114. Springer, 2001. 2

[22] H. P. Lensch, J. Kautz, M. Goesele, W. Heidrich, and H.-P. Seidel. Image-based reconstruction of spatial appearance and geometric detail. *ACM Transactions on Graphics (TOG)*, 22(2):234–257, 2003. 2

[23] T.-M. Li, M. Aittala, F. Durand, and J. Lehtinen. Differentiable monte carlo ray tracing through edge sampling. *ACM Transactions on Graphics (TOG)*, 37(6):1–11, 2018. 2

[24] Z. Li, M. Shafiei, R. Ramamoorthi, K. Sunkavalli, and M. Chandraker. Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and svbrdf from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2475–2484, 2020. 2

[25] Z. Li, Z. Xu, R. Ramamoorthi, K. Sunkavalli, and M. Chandraker. Learning to reconstruct shape and spatially-varying reflectance from a single image. *ACM Transactions on Graphics (TOG)*, 37(6):1–11, 2018. 2

[26] L. Liu, J. Gu, K. Z. Lin, T.-S. Chua, and C. Theobalt. Neural sparse voxel fields. *arXiv preprint arXiv:2007.11571*, 2020. 2

[27] S. Lombardi and K. Nishino. Radiometric scene decomposition: Scene reflectance, illumination, and geometry from rgb-d images. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 305–313. IEEE, 2016. 2

[28] M. McGuire. Computer graphics archive. https://casual-effects.com/data, July 2017. 8

[29] M. Meilland, C. Barat, and A. Comport. 3d high dynamic range dense visual slam and its application to real-time object re-lighting. In *2013 IEEE International Symposium*

*on Mixed and Augmented Reality (ISMAR)*, pages 143–152. IEEE, 2013. 2

[30] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 2

[31] G. Nam, J. H. Lee, D. Gutierrez, and M. H. Kim. Practical svbrdf acquisition of 3d objects with unstructured flash photography. *ACM Transactions on Graphics (TOG)*, 37(6):1–12, 2018. 2

[32] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohli, J. Shotton, S. Hodges, and A. Fitzgibbon. KinectFusion: Real-Time Dense Surface Mapping and Tracking. In *Proceedings of the International Symposium on Mixed and Augmented Reality (ISMAR)*, 2011. 3

[33] M. Nimier-David, D. Vicini, T. Zeltner, and W. Jakob. Mitsuba 2: A retargetable forward and inverse renderer. *ACM Transactions on Graphics (TOG)*, 38(6):1–17, 2019. 2

[34] S. G. Parker, J. Bigler, A. Dietrich, H. Friedrich, J. Hoberock, D. Luebke, D. McAllister, M. McGuire, K. Morley, A. Robison, and M. Stich. Optix: A general purpose ray tracing engine. *ACM Trans. Graph.*, 29(4):66:1–66:13, July 2010. 6, 7

[35] G. Patow and X. Pueyo. A survey of inverse rendering problems. In *Computer graphics forum*, volume 22, pages 663–687. Wiley Online Library, 2003. 2

[36] S. Peng, B. Haefner, Y. Quéau, and D. Cremers. Depth super-resolution meets uncalibrated photometric stereo. In *International Conference on Computer Vision Workshops (ICCVW)*, 2017. 2

[37] M. Pharr, W. Jakob, and G. Humphreys. *Physically based rendering: From theory to implementation*. Morgan Kaufmann, 2016. 2, 4, 5

[38] R. Ramamoorthi and P. Hanrahan. A signal-processing framework for inverse rendering. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 117–128, 2001. 2

[39] J. Riviere, P. Peers, and A. Ghosh. Mobile surface reflectometry. In *Computer Graphics Forum*, volume 35, pages 191–202. Wiley Online Library, 2016. 3

[40] L. Sang, B. Haefner, and D. Cremers. Inferring super-resolution depth from a moving light-source enhanced rgb-d sensor: A variational approach. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, Colorado, USA, March 2020. 2

[41] C. Schlick. An inexpensive brdf model for physically-based rendering. In *Computer graphics forum*, volume 13, pages 233–246. Wiley Online Library, 1994. 3

[42] C. Schmitt, S. Donne, G. Riegler, V. Koltun, and A. Geiger. On joint estimation of pose, geometry and svbrdf from a handheld scanner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[43] S. Sengupta, J. Gu, K. Kim, G. Liu, D. W. Jacobs, and J. Kautz. Neural inverse rendering of an indoor scene from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8598–8607, 2019. 2

[44] S. A. Shafer. Using color to separate reflection components. *Color Research & Application*, 10(4):210–218, 1985. 3

[45] J. Shi, Y. Dong, H. Su, and S. X. Yu. Learning non-lambertian object intrinsics across shapenet categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1685–1694, 2017. 2

[46] B. Smith. Geometrical shadowing of a random rough surface. *IEEE transactions on antennas and propagation*, 15(5):668–671, 1967. 3

[47] The Stanford 3D Scanning Repository. http://graphics.stanford.edu/data/3Dscanrep/. 8

[48] J. Straub, T. Whelan, L. Ma, Y. Chen, E. Wijmans, S. Green, J. J. Engel, R. Mur-Artal, C. Ren, S. Verma, A. Clarkson, M. Yan, B. Budge, Y. Yan, X. Pan, J. Yon, Y. Zou, K. Leon, N. Carter, J. Briales, T. Gillingham, E. Mueggler, L. Pesqueira, M. Savva, D. Batra, H. M. Strasdat, R. D. Nardi, M. Goesele, S. Lovegrove, and R. Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 6, 7, 8

[49] K. E. Torrance and E. M. Sparrow. Theory for off-specular reflection from roughened surfaces. *Josa*, 57(9):1105–1114, 1967. 3

[50] T. Trowbridge and K. P. Reitz. Average irregularity representation of a rough surface for ray reflection. *JOSA*, 65(5):531–536, 1975. 3

[51] B. Walter, S. R. Marschner, H. Li, and K. E. Torrance. Microfacet models for refraction through rough surfaces. *Rendering techniques*, 2007:18th, 2007. 3

[52] Y. Yu, P. Debevec, J. Malik, and T. Hawkins. Inverse global illumination: Recovering reflectance models of real scenes from photographs. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 215–224, 1999. 2, 4, 5, 8

[53] Y. Yu and J. Malik. Recovering photometric properties of architectural scenes from photographs. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, pages 207–217, 1998. 2

[54] Y. Yu, A. Meka, M. Elgharib, H.-P. Seidel, C. Theobalt, and W. A. Smith. Self-supervised outdoor scene relighting. In *European Conference on Computer Vision*, pages 84–101. Springer, 2020. 2

# Part III

# Conclusion and Outlook

# Chapter 9

# Summary

This thesis proposes novel algorithms to solve *physically-based inverse problems for high-quality 3D reconstruction*. Our methodology involved upsampling low-resolution (LR) depth maps from a commodity RGB-D sensor to the scale of the companion color image using a unique combination of Shape-from-Shading (SfS) and depth super-resolution (SR), resulting in high-quality, detail preserving depth maps. We then introduced a robust initialization and optimization solver to address the complex problem of uncalibrated photometric stereo (UPS) under natural illumination, leading to superior results compared to its state-of-the-art. Furthermore, we promoted an innovative combination of calibrated photometric stereo (CPS) with active contour segmentation, which eliminates the need for a segmentation mask, thus removing one of the fundamental assumptions of photometric stereo (PS) methods. Finally, we developed a new approach to estimating bidirectional reflectance distribution function (BRDF) parameters in scenes of significant scale. In the following, we summarize the core publications presented in Part II of this dissertation in terms of their novelties and key contributions.

**Single-Shot Depth Super-Resolution from Shading.** The ill-posed problems of SfS and depth SR have been effectively addressed in Chapter 5 using a single calibrated RGB-D image pair, thanks to a novel variational formulation that combines both problems. The low-frequency geometric information from the depth sensor helps disambiguate SfS, while high-frequency information in the color image helps to disambiguate depth SR. The depth SR is guided by a minimal surface regularizer in cases where no shape clues or RGB information is available, such as in the presence of holes in the depth image or under-/oversaturation in the intensity image. Furthermore, a novel piecewise constant albedo regularization term facilitates the robust recovery of crisp depth maps with unprecedented detail, as all smooth shading information is explained geometrically. This approach is suitable for a broad range of applications that require high-resolution geometric information and are based on RGB-D data that depicts a diffuse scene.

**Uncalibrated Photometric Stereo under General Lighting.** In Chapter 6, a novel variational paradigm has been presented for robustly recovering state-of-the-art depth maps from a set of UPS images under natural illumination. The recovery of high-quality depth maps from a set of UPS images has been achieved through the use of a robust depth initialization scheme based on minimal surfaces [167] and a perspective projection model that avoids ambiguities and non-integrabilities in the approach. To mitigate inaccuracies in the image formation model, Cauchy's robust M-estimator is utilized, which accounts for noise and quantization, lens aberrations, and non-local illumination artifacts such as specularities, inter-reflections and cast shadows. The resulting solution of the variational model yields state-of-the-art geometry estimates that are on average $2$–$3\times$ times better than the current state-of-the-art methods for UPS under general illumination.

**Simultaneous Photometric Stereo and Masking.** The methodology presented in Chapter 7 eliminates one of the fundamental assumptions in PS. The combination of active contour segmentation with a differentiable PS approach eliminates the need for the tedious preprocessing step to generate a segmentation mask. This significantly streamlines PS approaches and enables the generation of 3D objects from a set of PS images. To make this possible, the masking procedure is not solely based on 2D image information, *i.e.*, image brightness, but on 3D information by incorporating the image formation model, making the approach robust against strong intensity variations in the images. This is especially important in the case of PS techniques due to changing illumination and its resulting cast and self shadows. Unlike time-consuming manual or error-prone semi-automated approaches, this method results in a fast and robust simultaneous estimation of a shape and its silhouette.

**Recovering Reflectance and Shading From HDR Imagery.** In Chapter 8, we proposed a new framework for estimating BRDF parameters for all objects in large-scale scenes based on high dynamic range (HDR) data. First, we introduced a novel running median computation to estimate an approximate lit diffuse HDR texture of the scene. This allowed us to estimate the diffuse and non-diffuse characteristics of the dichromatic BRDF employed. For each surface point, we computed the albedo using a novel raytracing paradigm, effectively estimating each point's environment map. Next, we estimated the non-diffuse parameters per object via an innovative inverse rendering scheme that exploited the structure of the rendering equation and BRDF model. This involved performing a grid search w.r.t. the non-linear parameters of the BRDF and a nested closed-form least squares solution w.r.t. the linear parameters. Our results are of unprecedented detail and quality compared to related methods, making it possible to faithfully render the captured scene under novel illumination with virtually inserted objects.

# Chapter 10

# Future Research

Notwithstanding the advances made by this thesis' research, there exist certain constraints that can be utilized to steer future research endeavors. A few of these pursuits constitute contemporary open research questions in the realm of computer vision or computer graphics, while others encompass unambiguous expansions of the investigations expounded upon in this dissertation.

**RGB-D Cameras.** In the realm of depth sensing RGB devices, there persist obstacles that necessitate resolution. Since the advent of the *Microsoft Kinect V1* in 2010, headway has been made in the robustness and precision of depth measurements, augmentation of sensor resolution, resistance to water, physical dimension, employment of global shutter, inertial measurement unit (IMU), high frames per second (FPS), and other aspects. Present-day pinnacles in this field comprise the *MS Azure Kinect DK* and the *Intel RealSense* product line. Despite these advancements, state-of-the-art sensors continue to be susceptible to the very limitations that plagued the *Microsoft Kinect V1*.
All devices in this domain rely on a multi-sensor stereo system, whereby depth is typically computed using a stereo configuration, which can result in holes and missing geometry. Additionally, since the RGB and depth cameras are separate sensors, a post-processing step is required to align the RGB with the depth image, and this is heavily contingent upon the accuracy of the sensor's internal calibration, including intrinsic and extrinsic parameters as well as shutter synchronization between the RGB and depth sensors. Only a few research directions have been suggested to mitigate these challenges [85, 215], which involves using beam splitter devices.
Furthermore, despite a general improvement in the resolution of both sensors, there persists a gap between the resolution of the RGB and depth sensors, and discretization artifacts in the depth data still result in the depth super-resolution (SR) problem being relevant in the context of RGB-D rigs.

**Inverse Rendering.** In the context of inverse rendering, retrieval of shape, material, light, and camera properties remains a vibrant area of research. The complexity of this problem is attributed to the presence of non-convex, non-differentiable, and ill-posed cost functions. While some approaches have made headway in tackling these challenges [18, 126, 162, 163, 231, 249], their usage is limited due to their high demand for (graphics processing unit (GPU)) memory and long runtime in even the simplest of cases, rendering them impractical for larger-scale scenarios. This is mainly due to the use of general-purpose optimization schemes, particularly stochastic gradient descent (SGD) based on automatic differentiation.
As evidenced by our findings in Chapter 8, an optimization scheme tailored specifically to the inverse rendering problem has the potential to yield more accurate results and faster runtimes, making it a promising avenue for future research. In this regard, convex relaxation methods, such as functional lifting approaches [77, 155], could also be explored. While these methods are generally memory-intensive, recently proposed scalable approaches have shown promise in addressing this challenge [10, 11]. Nevertheless, the applicability of such methods to inverse rendering problems remains unexplored.
Furthermore, it is commonly assumed that one or more scene assets are known upfront in inverse rendering, such as in the case of NeRFs for image synthesis [152], geometry reconstruction [231, 249], or light-field and material estimation [248]. These assumptions include calibrated images [152, 231, 248, 249] and even known geometry [248]. Recently, approaches have been developed to mitigate the calibration assumption [31, 124, 202, 232]. However, these approaches require user interaction [31] or are only applicable in the context of image synthesis [124, 202, 232], meaning that they do not specifically estimate geometry, material, or lighting. As a result, solving the inverse rendering problem in terms of scene and camera assets remains an open research challenge.

**Photometric Stereo.** While the problem of photometric stereo (PS) has been studied since Woodham's pioneering work [238], recent trends have focused on deep learning approaches to improve the robustness and accuracy of the method. Although these approaches can handle non-diffuse reflectance well, they still only consider limited setups, such as (calibrated) directional lighting, only estimating potentially non-integrable normals, or requiring large amounts of training data [43, 44, 45, 87, 103, 129, 199, 200].
Therefore, shape estimation in the context of deep PS remains an interesting research alley. While a few recent works, such as those on near-light illumination [84] or global lighting contexts [97, 98], have tackled different scenarios, there is still much to be explored in terms of unsupervised learning-based approaches that optimize for shape under various illumination setups, which have the potential to further advance the field beyond the capabilities of model-based methods.

Furthermore, an area that remains underexplored in PS is the identification of ill- and well-posed setups. A systematic investigation of this issue could offer valuable theoretical insights into the underlying assumptions required to achieve robust and accurate shape recovery.

**BRDF Parameter Estimation.** In the realm of material estimation, most existing methods presuppose an isotropic, non-metallic, and opaque bidirectional reflectance distribution function (BRDF). These employ the widely adopted microfacet model [51, 221] with the Trowbridge-Reitz (GGX) distribution [222, 230], or employ artistic models such as the Disney BRDF [38].
While this setup is already fairly complex, it still imposes limitations on the range of applications of these methods. Specifically, it is currently not possible to recover transparent materials, such as glass or smoke, from a set of real-world images. Thus, it would be interesting to explore incorporating the bidirectional transmittance distribution function (BTDF) to overcome these limitations.
The representation of BRDF functions using neural networks is an open problem. While there are possibilities for deploying multilayer perceptrons (MLPs) to represent shape [231, 249], or light-fields [248, 254], no literature exists on how to formulate a similar approach for BRDFs. Currently, a parametric BRDF model is often employed, as described above, where the BRDF parameters are represented using MLPs. However, an interesting approach could be to represent the full BRDF as a ”black-box” MLP, as this could potentially mitigate the limitations on anisotropy or dielectricity.

To propose potential future research directions, the following four paragraphs outline possible extensions to the core publications presented in this thesis.

**Single-Shot Depth Super-Resolution from Shading (Chapter 5).** Although our approach can generate high-quality SR depth maps, it is contingent on several assumptions. Specifically, we rely on the underlying material of the object to be a piecewise-constant albedo. However, this assumption may not hold true for all scenarios, and therefore, a possible extension to our approach could involve the adoption of a smooth albedo assumption. This could be achieved by exploring the differences in albedo and shading variations. For instance, under achromatic lighting conditions, shading variations tend to be similar across all three RGB channels, whereas albedo changes may differ. This approach has been investigated in several previous works by the research group of Steven Zucker [16, 24, 91, 122] using the hue channel in the hue, saturation, value (HSV) color space. Another possible avenue for future research would be to incorporate non-diffuse behavior in our approach, which has already been explored in a

similar context by incorporating the infrared (IR) image from the RGB-D sensor [62].
The current runtime figures are based on an inefficient MATLAB implementation, although some of the steps are performed using a CUDA framework. Nevertheless, a custom GPU implementation has the potential to significantly improve the runtime performance of our approach, bringing it closer to real-time operation. This would be particularly advantageous in a multi-view scenario. Additionally, estimated quantities from the preceding frame could be utilized to initialize the new frame, and some form of temporal regularization, similar to that employed in [216], could be introduced.
Furthermore, this approach could be extended to develop pipelines similar to those in [140, 263], which integrate multiple RGB-D frames to create an signed distance function (SDF) volume of a larger scene.

**Uncalibrated Photometric Stereo under General Lighting (Chapter 6).** The novelty of this approach lies in the use of a robust optimization solver, as well as minimal surface regularization, which is essential in achieving high-quality depth maps. However, the process of hyperparameter tuning is required to obtain a suitable initialization. To simplify this procedure, one potential avenue for future research would be to implement the uncalibrated photometric stereo (UPS) solver in a coarse-to-fine scheme. It is well-known that such an approach is less sensitive to initialization, and therefore could potentially reduce the amount of manual tuning required.
Moreover, the current approach assumes a purely diffuse scene and treats specularities as outliers. To produce a more photorealistic reconstruction of the scene, it is necessary to take into account the non-diffuse part of the BRDF as well. This could be achieved either as a postprocessing step or to incorporate the specular component directly into the reconstruction process by utilizing neural approaches, as has already been achieved in previous works described earlier.

**Simultaneous Photometric Stereo and Masking (Chapter 7).** Although the presented methodology is convenient in that PS can be implemented without the need for a mask, it has only been tested under specific conditions, such as calibrated photometric stereo (CPS) with directional lighting and orthographic projection. A promising direction for future research would be to expand this methodology to encompass a wider range of PS approaches, both calibrated and uncalibrated, to create a versatile setup that can be applied to various PS tasks.
As with the approach described above, tuning a hyperparameter is necessary to obtain high-quality depth maps and segmentation masks. To streamline this process, one possible avenue for future research would be to develop a more efficient implementation that provides faster feedback after parameter tuning. Alternatively, incorporating segmen-

tation approaches that are less sensitive to hyperparameters, such as neural networks, could be explored. Recently, there have been promising advancements in this area, such as the deployment of PS without a mask and on multiple objects simultaneously [97].

**Recovering Reflectance and Shading From HDR Imagery (Chapter 8).** Given the reliance of the diffuse component on the quality of the median texture, specularities baked into the texture may propagate to the albedo when specular highlights are present in more than $50\%$ of the captures. To address this issue, prospective investigations could explore either a more accurate estimation of the lit diffuse texture that avoids baked-in specularities or a more sophisticated approach for estimating the albedo in large-scale scenes.
In addition, future studies could consider exploring the relaxation of the assumption of a constant non-diffuse material per object, given that most objects are composed of multiple non-diffuse materials. One possible avenue is to investigate a neural approach that learns the segmentation of the non-diffuse material without explicitly estimating the non-diffuse BRDF parameters, which can then be incorporated into our framework. Alternatively, a more elaborate procedure could be employed to estimate a non-diffuse spatially varying BRDF (SVBRDF) per object.
The approach presented in this study utilizes a single bounce for image rendering, which may lead to a degradation in the results due to inter-reflections. To mitigate this issue, exploring the incorporation of multiple bounces in the solver to improve the accuracy of the rendered images is a promising direction.
Exploring methods to recover a wider range of materials is also of great interest for future research. This can be accomplished by either incorporating more parameters of the Disney BRDF or by investigating alternative BRDF models.
Additionally, the reconstruction of translucent materials, such as glass, presents another promising avenue for research. One potential approach is to additionally optimize physically realistic BTDF models, or to explore the Disney bidirectional scattering distribution function (BSDF) [37][1].

[1] A BSDF is a combination of BRDF and BTDF, and enables the rendering of both opaque and translucent materials.

# Part IV

# Appendix

# Fight ill-posedness with ill-posedness: Single-shot variational depth super-resolution from shading Supplementary material

Bjoern Haefner Yvain Quéau Thomas Möllenhoff Daniel Cremers
Department of Informatics, Technical University of Munich, Germany
{bjoern.haefner,yvain.queau,thomas.moellenhoff,cremers}@tum.de

## 1. Additional real-world experiments

We ran our algorithm against two publicly available datasets [1, 3] to further demonstrate the effectiveness of our method.
Both datasets offer RGB-D frames, whereas the RGB images have resolutions of $1280 \times 1024$ px$^2$, $1296 \times 968$ px$^2$ and $640 \times 480$ px$^2$, respectively and the depth images come with a resolution of $640 \times 480$ px$^2$. Additionally, the corresponding multi-view reconstructions based on each of the methods described in [2, 4] are provided.
Figures 1, 2, 3, 4 show that our method provides good depth estimates on each of the additional datasets. Even in the case of cast- or self-shadows we are able to recover fine details of the depth without inducing too strong bias from the companion color image, see Figure 2 the cast-shadow of the camera or Figure 4 the self-shadows. Our method also seems to be robust to more complex lighting, see Figure 3 that the upper-right area of the RGB image is much darker compared to the well illuminated lower-left area of the image.

## References

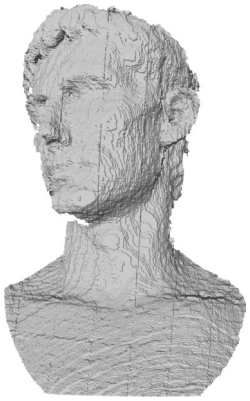
[1] R. Maier, K. Kim, D. Cremers, J. Kautz, and M. Nießner. Intrinsic3D Dataset. http://vision.in.tum.de/data/datasets/intrinsic3d, 2017. 1, 5

[2] R. Maier, K. Kim, D. Cremers, J. Kautz, and M. Nießner. Intrinsic3d: High-quality 3D reconstruction by joint appearance and geometry optimization with spatially-varying lighting. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 1, 5

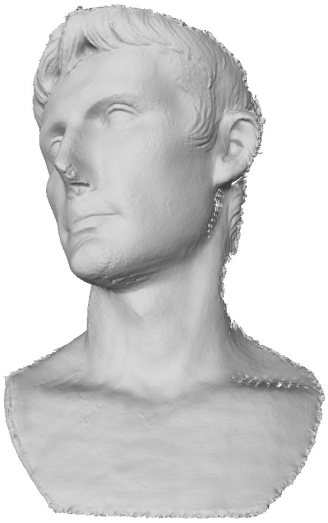
[3] M. Zollhöfer, A. Dai, M. Innman, C. Wu, M. Stamminger, C. Theobalt, and M. Nießner. Shading-based Refinement on Volumetric Signed Distance Functions. http://graphics.stanford.edu/projects/vsfs/, 2015. 1, 2, 3, 4

[4] M. Zollhöfer, A. Dai, M. Innman, C. Wu, M. Stamminger, C. Theobalt, and M. Nießner. Shading-based refinement on volumetric signed distance functions. *ACM Transactions on Graphics*, 34(4):96:1–96:14, 2015. 1, 2, 3, 4

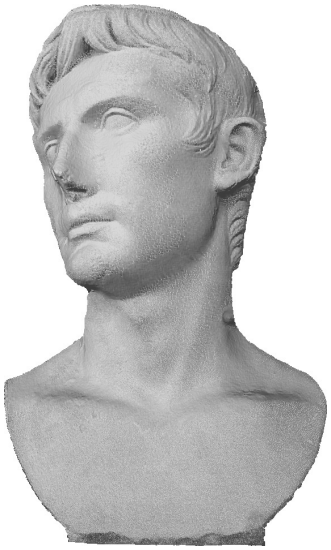
(a) RGB input



(b) Depth input

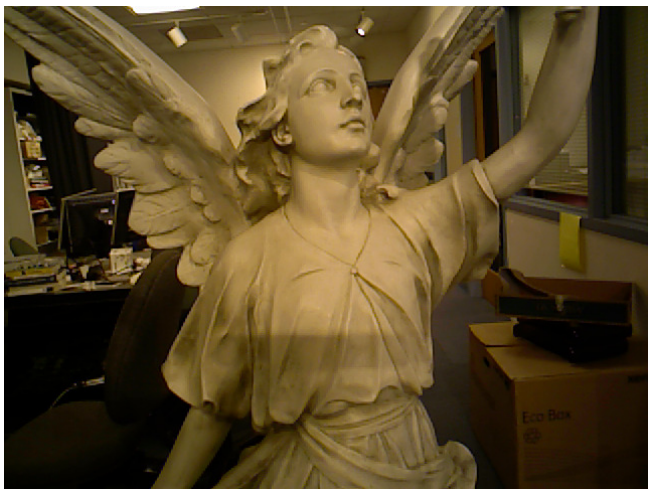


(c) Result of the multi-view approach [4]



(d) Our result using a single RGB-D frame

Figure 1: Augustus dataset of [3]



(a) RGB input

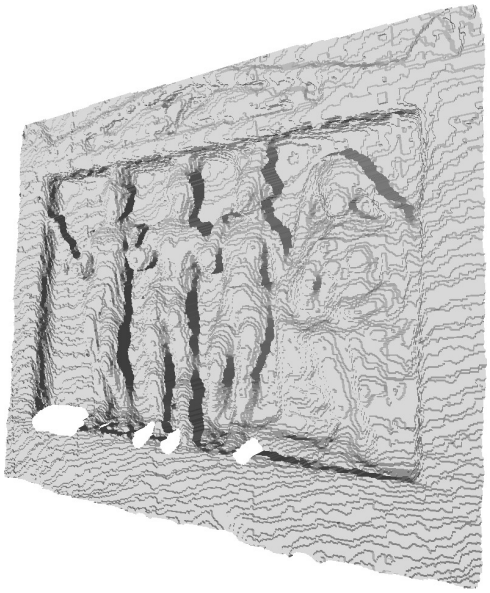
(b) Depth input

(c) Result of the multi-view approach [4]

(d) Our result using a single RGB-D frame

Figure 2: Lucy dataset of [3]


(a) RGB input



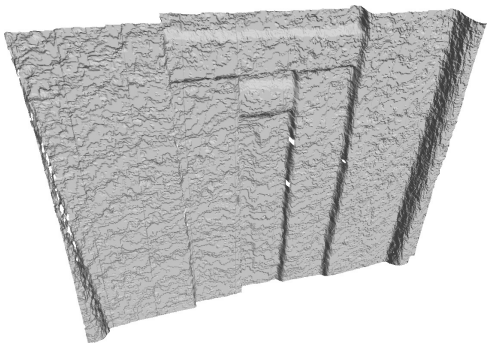
(b) Depth input

(c) Result of the multi-view approach [4]

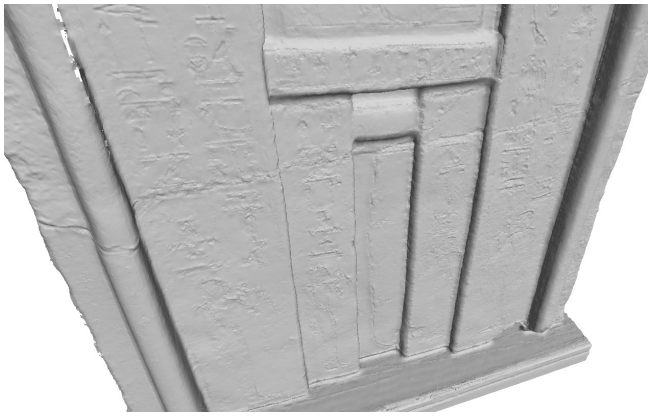

(d) Our result using a single RGB-D frame

Figure 3: Relief dataset of [3]

(a) RGB input



(b) Depth input



(c) Result of the multi-view approach [2]

(d) Our result using a single RGB-D frame

Figure 4: Gate dataset of [1]

# Variational Uncalibrated Photometric Stereo under General Lighting Supplementary Material

Bjoern Haefner*, [1] Zhenzhang Ye*, [1] Maolin Gao[2] Tao Wu[1] Yvain Quéau[3] Daniel Cremers[1]
[1]Technical University of Munich [2]Artisense [3]GREYC, UMR CNRS 6072
{bjoern.haefner, zz.ye, tao.wu, cremers}@tum.de maolin@artisense.ai yvain.queau@ensicaen.fr

## 1. Further Details on Synthetic Experiments

To provide further insights on the synthetic experiments (in Section 6.1), we visualize the environment lightings $\ell^i$, $i = 1 \dots 25$, used to render each image. Figure 1 shows all 25 environment maps[1]. The impact of each incident lighting $\ell^i$, $i = 1 \dots 25$, is illustrated in Figure 2 showing the Joyful Yell with a White ($\rho \equiv 1$) albedo. Thus, color changes in the images are caused by lighting only, as depicted in model (1) and (7) in the main paper.

Table 1 shows the mean angular error (MAE) of each dataset on the state-of-the-art approaches [1, 2, 3] and our proposed methodology. It can be seen that our approach consistently overcomes [1, 2, 3] by a factor of 2–3. Only the Pattern albedo seems to bias the resulting depth negatively, yet even in this case our approach estimates the geometry more faithfully than the current state-of-the-art.

Two more qualitative results on synthetic data are shown in Figure 3. While [1] gives more meaningful results on Armadillo with Constant albedo, depth deteriorates strongly on Lucy with Hippie albedo. Methods of [2, 3] both result in rather flattened shapes (cf. Lucy). Most accurate results are achieved using the proposed method where fine geometric details, as well as non flattened depth estimates are shown.

Additional to the depth results, Figure 4 shows estimated lightings and albedos along with the ground truths. Although lighting estimates show less shadowed areas and seem brighter compared to ground truths, this does not seem to affect reflectance estimations much. The estimated albedos are satisfactory, although some shading information is slightly visible.

The initialization is indeed crucial for the whole algorithm. Here, we show two different non-trivial initializations for our algorithm in Table 1: 1) Hemisphere, we first compute the circumscribed sphere for the 3D points of ground truth. The projection of each point onto this sphere is considered as initialization; 2) Initialization by [2], we simply refine the result from [2] by our algorithm. In Figure 5, we show visualized results. In certain special cases, the initialization from [2] is slightly better. However, our minimal surface strategy is stable for all cases, and our algorithm improves the results from [2]) in most cases.

## 2. Further Details on Real-World Results

Supplementary to the real-world experiments (in Section 6.2), Figures 6 and 7 show alternative viewpoints of the real-world results. The estimated albedos, which are mapped onto the surfaces, appear satisfactory. Correspondingly, we also show the estimated albedos and lightings. In view of the multiplicative ambiguity between lightings and albedos, all visualized albedos are normalized to have maximum value 1.

## References

[1] Paolo Favaro and Thoma Papadhimitri. A closed-form solution to uncalibrated photometric stereo via diffuse maxima. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 821–828, 2012. 1, 4, 5

[2] Zhipeng Mo, Boxin Shi, Feng Lu, Sai-Kit Yeung, and Yasuyuki Matsushita. Uncalibrated photometric stereo under natural illumination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2936–2945, 2018. 1, 4, 5, 7

[3] Songyou Peng, Bjoern Haefner, Yvain Quéau, and Daniel Cremers. Depth super-resolution meets uncalibrated photometric stereo. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, pages 2961–2968, 2017. 1, 4, 5

*Authors contributed equally.

[1]All environment maps were downloaded from http://www.hdrlabs.com/sibl/archive.html

Figure 1. All environment maps $\ell^i$ (360° view) used throughout the synthetic evaluation.

Figure 2. Illustration of the input data. The Joyful Yell dataset with White albedo to show the impact of the different environment maps used throughout the synthetic experimental validation.

| Dataset | | [1] | [3] | [2] | Our approach with different initializations | | |
|---|---|---|---|---|---|---|---|
| Shape | Albedo | | | | Hemisphere | Using [2] | Minimal surface (Sec. 5.1) |
| Armadillo | Bars | 26.22 | 27.84 | 36.91 | 79.54 | 20.08 | **16.78** |
| | Constant | 25.84 | 26.64 | 36.87 | 83.01 | 18.81 | **13.97** |
| | Ebsd | 25.34 | 26.88 | 27.80 | 82.53 | 15.99 | **14.26** |
| | Hippie | 28.21 | 27.30 | 25.82 | 79.12 | **12.56** | 14.52 |
| | Lena | 27.07 | 27.33 | 28.36 | 84.24 | 17.79 | **14.78** |
| | Pattern | 45.87 | 26.82 | 24.01 | 82.59 | 19.39 | **19.06** |
| | Rectcircle | 26.97 | 26.71 | 36.23 | 80.68 | 19.64 | **14.06** |
| | Voronoi | 25.62 | 26.91 | 50.70 | 79.65 | 55.29 | **14.07** |
| | White | 26.19 | 26.64 | 52.04 | 83.04 | 56.74 | **14.13** |
| Joyful Yell | Bars | 21.84 | 16.26 | 31.80 | 21.21 | 28.82 | **8.69** |
| | Constant | 23.95 | 14.93 | 33.47 | 16.85 | 29.31 | **5.96** |
| | Ebsd | 26.08 | 15.63 | 15.91 | 17.63 | 7.49 | **7.28** |
| | Hippie | 28.67 | 16.23 | 22.96 | 17.68 | **7.47** | 7.49 |
| | Lena | 21.33 | 16.33 | 19.70 | 20.11 | 13.16 | **9.21** |
| | Pattern | 26.07 | 18.76 | 26.67 | 18.76 | 21.03 | **16.97** |
| | Rectcircle | 35.27 | 15.19 | 52.41 | 16.27 | 61.77 | **7.34** |
| | Voronoi | 22.27 | 16.42 | 45.74 | 18.62 | 54.78 | **6.57** |
| | White | 27.12 | 14.32 | 33.06 | 17.70 | 28.99 | **6.20** |
| Lucy | Bars | 49.13 | 21.90 | 36.51 | 40.55 | 26.15 | **8.16** |
| | Constant | 54.98 | 19.89 | 36.57 | 41.00 | 25.74 | **8.71** |
| | Ebsd | 62.33 | 20.81 | 23.56 | 40.80 | 13.36 | **9.61** |
| | Hippie | 58.61 | 21.29 | 32.38 | 39.93 | 8.10 | **7.87** |
| | Lena | 64.01 | 22.24 | 30.93 | 40.16 | 19.14 | **9.56** |
| | Pattern | 48.83 | 22.25 | 32.68 | 40.11 | 20.56 | **17.78** |
| | Rectcircle | 24.68 | 20.99 | 43.13 | 41.17 | 10.01 | **8.98** |
| | Voronoi | 61.53 | 22.10 | 48.14 | 40.39 | 71.32 | **7.59** |
| | White | 64.43 | 19.33 | 44.76 | 41.54 | 72.45 | **8.76** |
| Thai Statue | Bars | 25.53 | 21.91 | 66.17 | 78.72 | 8.94 | **8.55** |
| | Constant | 27.20 | 18.91 | 38.47 | 81.14 | 24.26 | **9.58** |
| | Ebsd | 27.85 | 20.22 | 34.11 | 79.58 | 19.23 | **9.47** |
| | Hippie | 21.91 | 21.86 | 30.62 | 77.27 | 12.78 | **8.83** |
| | Lena | 33.53 | 19.66 | 34.00 | 79.43 | 19.55 | **9.19** |
| | Pattern | 26.77 | 22.06 | 28.81 | 83.92 | 16.69 | **15.27** |
| | Rectcircle | 29.36 | 19.92 | 43.86 | 81.88 | 79.88 | **8.84** |
| | Voronoi | 30.65 | 21.56 | 36.58 | 78.92 | 25.21 | **8.69** |
| | White | 28.02 | 18.64 | 37.31 | 81.54 | 24.94 | **9.16** |
| Median | | 27.16 | 21.14 | 34.06 | 59.41 | 19.86 | **9.17** |
| Mean | | 34.15 | 21.18 | 35.53 | 55.20 | 27.43 | **10.72** |

Table 1. Quantitative comparison between our method and other state-of-the-art methods on challenging synthetic datasets. The last three columns refer to the results with different initializations for our approach.

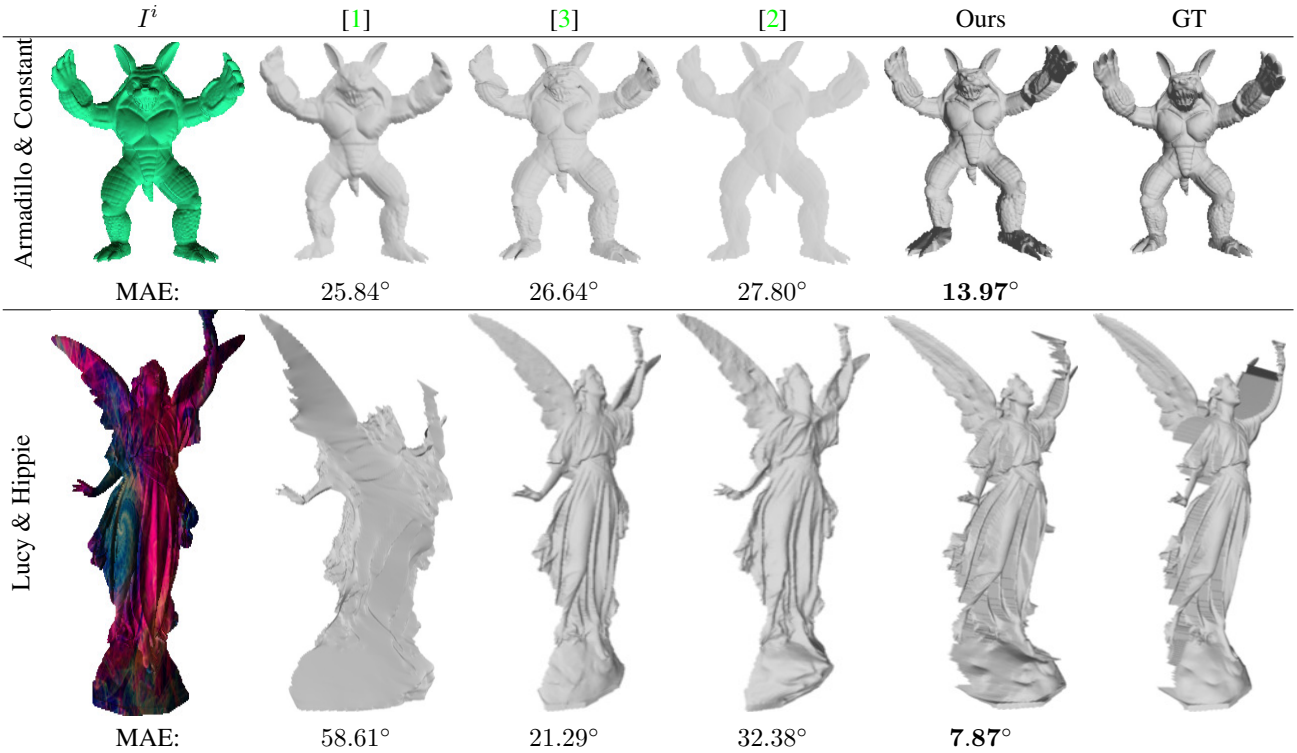



Figure 3. Results of state-of-the-art approaches and our approach on two out of the 36 synthetic datasets. Numbers show the mean angular error (MAE) in degrees.

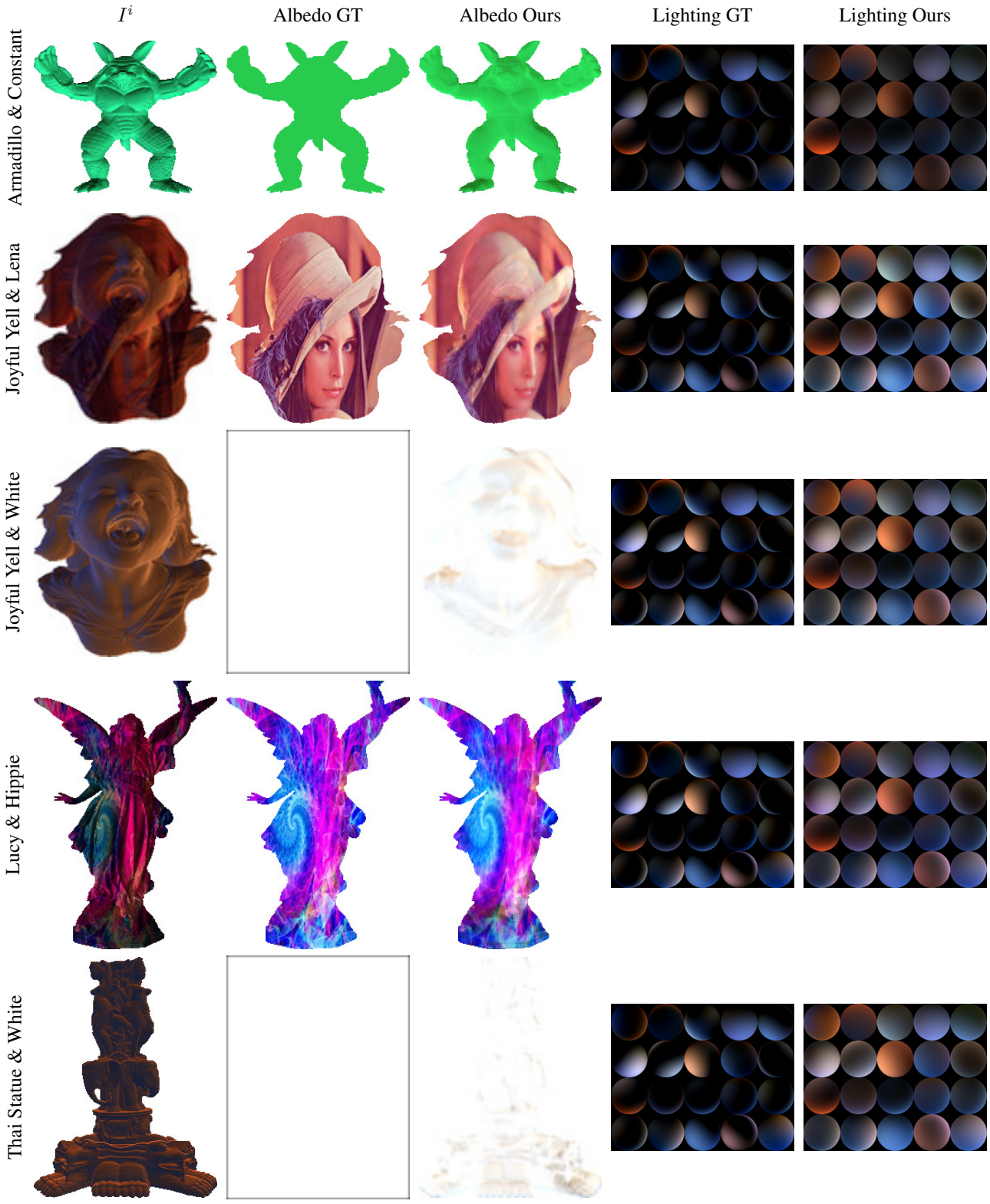



Figure 4. Our estimated albedos and lighting next to the ground truth. Lighting estimates show less shadowed areas and seem brighter compared to ground truth, yet this does not seem to affect reflectance and geometry estimation much, cf. Figure 7 in main paper and Figure 3 in the supplementary material. The estimated albedos are satisfactory, although some shading information is slightly visible.

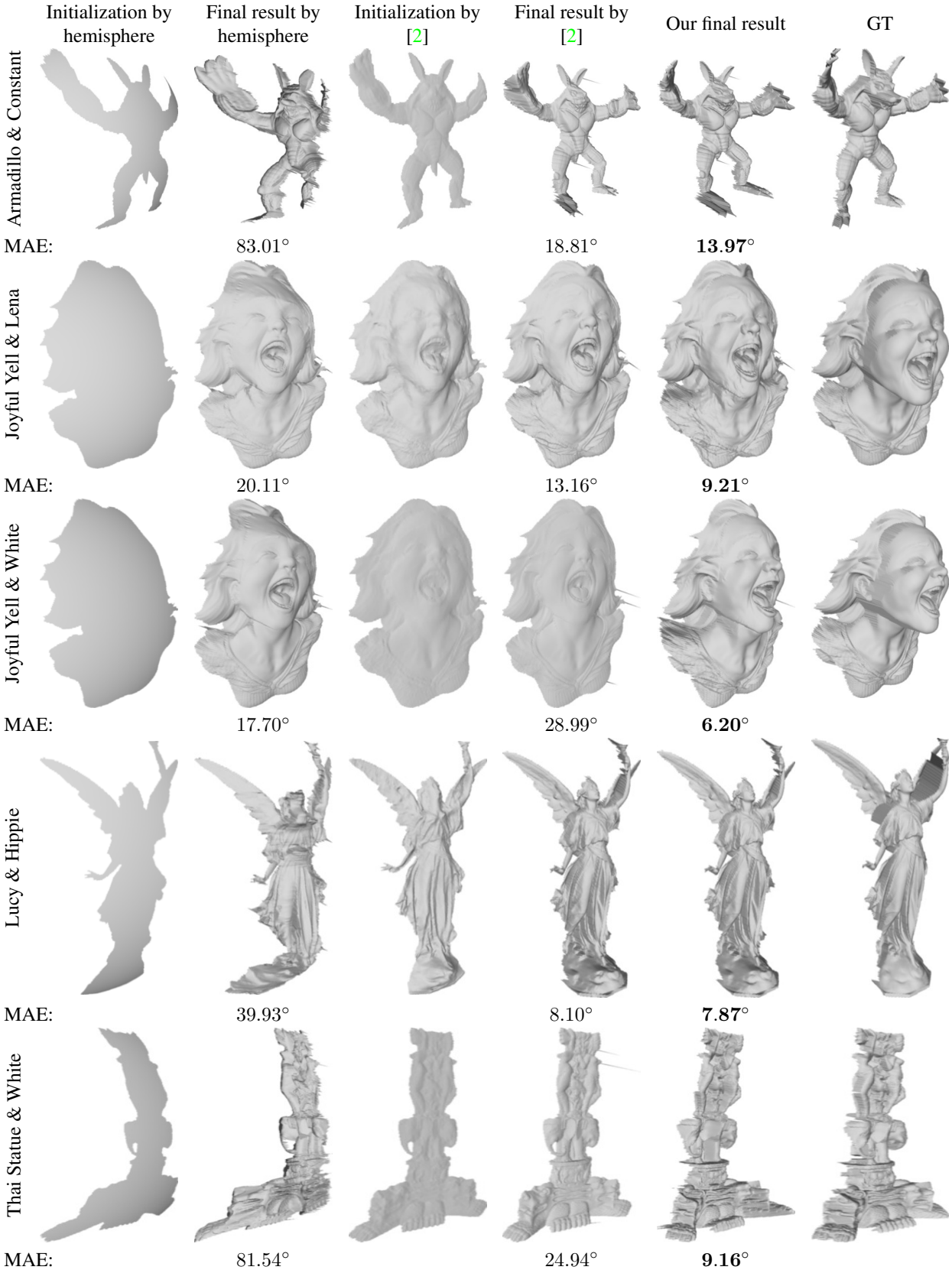



Figure 5. Our results compared those from two different initializations of our algorithm. Numbers show the mean angular error (MAE) in degrees. Though the initialization by [2] achieves comparable result to ground truth on "Lucy & Hippie" dataset, its performance is not stable across different datasets.

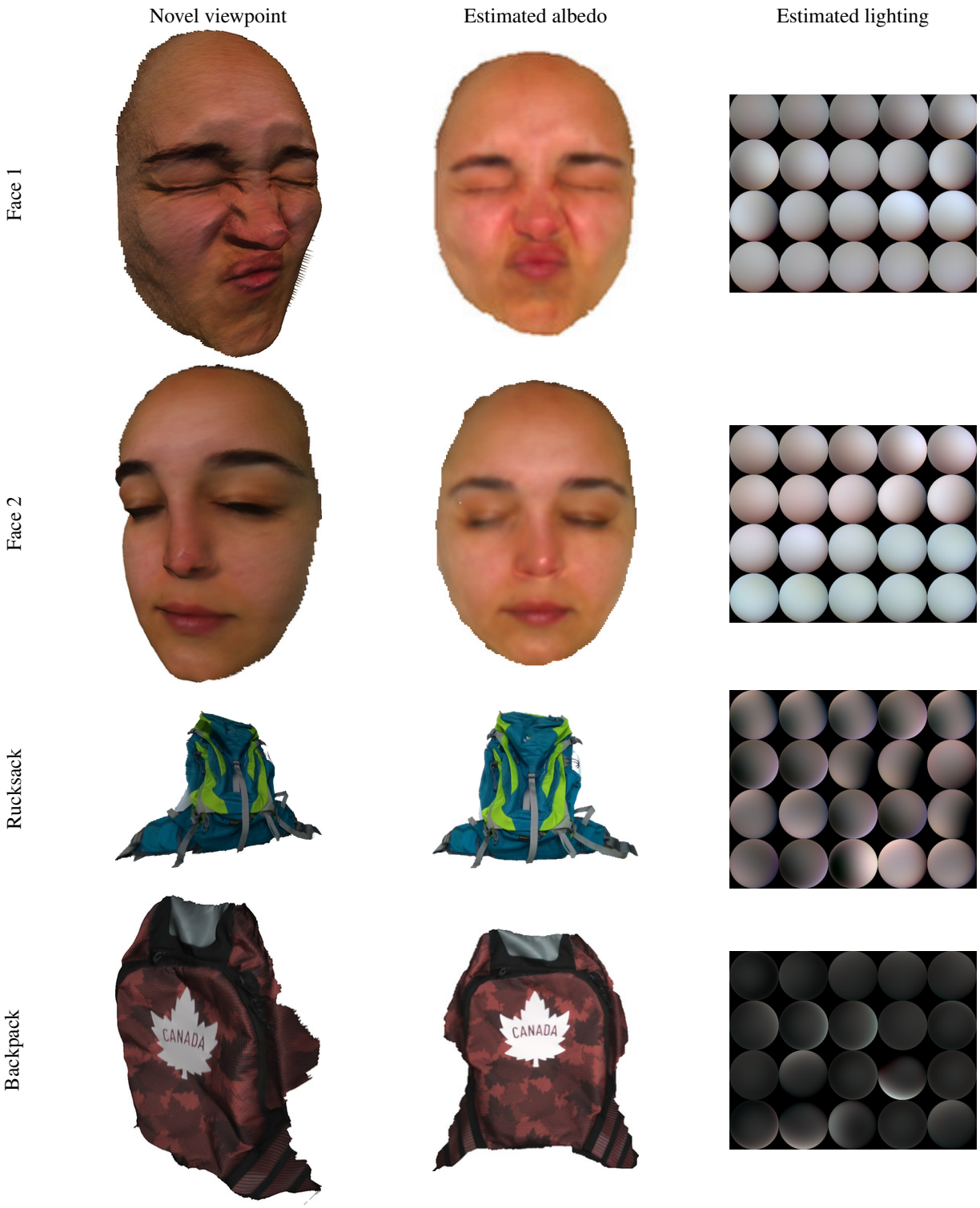



Figure 6. Real-world results: (left) estimated albedos mapped onto estimated surfaces rendered under a novel viewpoint, (middle) estimated albedos, (right) estimated lightings for all $M = 20$ input images.

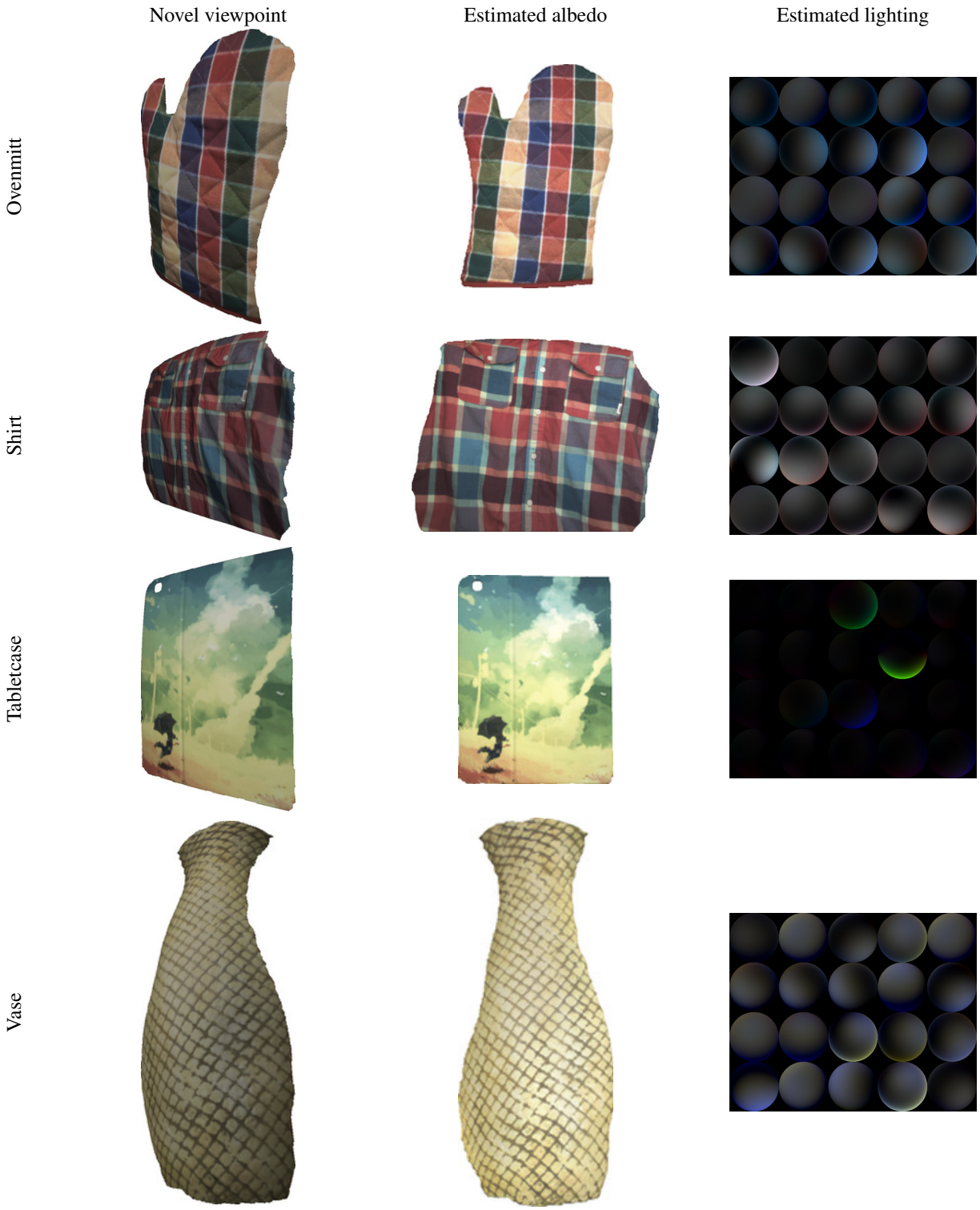



Figure 7. More real-world results: (left) estimated albedos mapped onto estimated surfaces rendered under a novel viewpoint, (middle) estimated albedos, (right) estimated lightings for all $M = 20$ input images.

# Supplementary Material of
# Recovering Real-World Reflectance Properties and Shading From HDR Imagery

Bjoern Haefner[1,2] Simon Green[2] Alan Oursland[2] Daniel Andersen[2]
Michael Goesele[2] Daniel Cremers[1] Richard Newcombe[2] Thomas Whelan[2]
[1]Technical University of Munich [2]Facebook Reality Labs Research
{bjoern.haefner, cremers}@tum.de,
{simongreen, ours, andersed, goesele, newcombe, twhelan}@fb.com

## A. Details on importance sampling (Section 3.4.2)

This section discusses the technical implementation details on an efficient sampling strategy to evaluate

$$I_{\text{nd}}\left(\boldsymbol{p};\varphi,\psi\right) := \int_{\mathcal{H}^2} f_{\text{nd}}\left(\boldsymbol{x},\omega,\omega_{\text{o}};\varphi,\psi\right) L\left(\boldsymbol{x},\omega\right) \left\langle\omega,\boldsymbol{n}\right\rangle \mathrm{d}\omega. \tag{1}$$

Importance sampling is a powerful tool to estimate the integral in (1) and we refer the interested reader to [9], Chapter 13 for the mathematical reasoning behind it. The Monte-Carlo estimator used for Eqn. (1) using the non-diffuse part of the simplified Disney BRDF is the finite sum of the form,

$$\mathcal{I}_{\text{nd}}(\boldsymbol{p};\varphi,\psi)=\frac{1}{N}\sum_{j=1}^{N}\frac{f_{\text{nd}}(\boldsymbol{x},\boldsymbol{\Omega}_j,\omega_{\text{o}};\varphi,\psi)L(\boldsymbol{x},\boldsymbol{\Omega}_j)\langle\boldsymbol{\Omega}_j,\boldsymbol{n}\rangle}{p(\boldsymbol{\Omega}_j)}, \tag{2}$$

where the random variables $\boldsymbol{\Omega}_j$ are samples drawn from the probability density function $p(\omega)$. We expect, i.e. given enough samples $N$,

$$\mathbb{E}\left[\mathcal{I}_{\text{nd}}\left(\boldsymbol{p};\varphi,\psi\right)\right] = I_{\text{nd}}\left(\boldsymbol{p};\varphi,\psi\right), \quad \forall \boldsymbol{p},\varphi,\psi. \tag{3}$$

The probability density function used in our approach is

$$p\left(\omega\right) = \frac{1}{2}\frac{\left|\langle\omega,\boldsymbol{n}\rangle\right|}{\pi} + \frac{1}{2}\frac{D\left(\varphi\right)\left|\langle h,\boldsymbol{n}\rangle\right|}{4\left|\langle\omega_{\text{o}},h\rangle\right|}. \tag{4}$$

To evaluate (2), we need to be able to sample random variables $\boldsymbol{\Omega}_j$ from $p(\omega)$ and we realize this the following way: Given the $j$-th observation of random variables following a uniform distribution over $[0,1)$, $X_0^j, X_1^j, X_2^j \sim \mathcal{U}(0,1)$, we calculate a sample of incident direction as

$$\boldsymbol{\Omega}_j = \begin{cases} T s_{\mathcal{H}^2}(X_1^j, X_2^j), & X_0^j < \frac{1}{2}, \\ R(\omega_{\text{o}}, T h_s(X_1^j, X_2^j)), & \text{else}, \end{cases} \tag{5}$$

where $R(\omega_{\text{o}},h) = 2\left\langle\omega_{\text{o}},h\right\rangle h - \omega_{\text{o}}$ resembles the reflection of $\omega_{\text{o}}$ on $h$, and $T := (\mathbf{t}_1,\mathbf{t}_2,\mathbf{t}_3) \in \mathbb{R}^{3\times3}$ is an orthonormal basis transform in the normal's coordinate system, aligning the north pole of $\mathcal{H}^2$ with the normal $\boldsymbol{n}$,

$$\mathbf{t}_1 = \mathbf{t}_2 \times \mathbf{t}_3 \tag{6}$$

$$\mathbf{t}_2 = \begin{cases} \frac{(-\boldsymbol{n}_y,\boldsymbol{n}_x,0)^\top}{\|(-\boldsymbol{n}_y,\boldsymbol{n}_x,0)\|}, & |\boldsymbol{n}_x| > |\boldsymbol{n}_z|, \\ \frac{(0,-\boldsymbol{n}_z,\boldsymbol{n}_y)^\top}{\|(0,-\boldsymbol{n}_z,\boldsymbol{n}_y)\|}, & \text{else}, \end{cases} \tag{7}$$

$$\mathbf{t}_3 = \boldsymbol{n}. \tag{8}$$

To sample the diffuse lobe of the BRDF (the case when $X_0^j < \frac{1}{2}$ in (5)), we generate random samples on the upper hemisphere $\mathcal{H}^2$ using $s_{\mathcal{H}^2} : [0,1)^2 \to \mathcal{H}^2$,

$$s_{\mathcal{H}^2}(x_1,x_2) = \begin{pmatrix} s_1 \\ s_2 \\ \sqrt{\max\left(0, 1-s_1^2-s_2^2\right)} \end{pmatrix}, \tag{9}$$

with $s_1 := \sqrt{x_1}\cos\left(2\pi x_2\right)$ and $s_2 := \sqrt{x_1}\sin\left(2\pi x_2\right)$. The non-diffuse lobe of the BRDF (the case when $X_0^j \geq \frac{1}{2}$ in (5)) is sampled using $h_s : [0,1)^2 \to \mathcal{H}^2$,

$$h_s(x_1,x_2) = \begin{pmatrix} \sin\left(\theta\right)\cos\left(2\pi x_1\right) \\ \sin\left(\theta\right)\sin\left(2\pi x_1\right) \\ \cos\left(\theta\right) \end{pmatrix}, \tag{10}$$

with $\theta := \cos^{-1}\left(\sqrt{\frac{1-x_2}{1+(\varphi^2-1)x_2}}\right)$.

## B. Details on capturing process (Section 4)

We perform two full scans of a room sized environment, where camera poses are recovered using SLAM [2, 7], geometry is reconstructed with [8]. In a post-processing step we fill larges holes manually or using Poisson reconstruction [4, 5] and repair any remaining issues automatically using [3]. Object segmentation is carried out in a manual step.

## C. Further quantitative results on albedo and shading estimation validation (Section 4.1)

Additional qualitative results of the albedo and shading estimation applied to real-world data sets are shown in Figure 1.

## D. Further quantitative results on specular appearance estimation validation (Section 4.2)

This section discusses **further quantitative results** of the specular appearance estimation for novel views. The main paper depicts quantitative results as well as a qualitative visualization of notable peaks in the corresponding graph on the "Office 1" sequence of [11]. For full insight, we show the results on the remaining sequences of the Replica datasets [11], cp. Figure 2 for insight in the "Office" sequences using the $L_2$ metric, Figure 3 for insight in the "Office" sequences using the FLIP evaluator [1], Figure 4 for insight in the "Room" sequences using the $L_2$ metric, and Figure 5 for insight in the "Room" sequences using the FLIP evaluator [1].
Figures 2 and 3 show results on the "Office" sequences of [11], they consist of 1293, 2117, 2459, and 2101 frames, which include 24, 38, 43, and 31 target frames, respectively, thus incorporating 1269, 2079, 2416, and 2070 novel, unseen viewpoints. The "Office 0", "Office 2", "Office 3", and "Office 4" sequences have their largest improvement and deterioration for the $L_2$ error at frames $(1264, 243)$, $(629, 910)$, $(1799, 2319)$, and $(1899, 1903)$, respectively and are visualized for qualitative inspection in Figure 2. The same sequences have their largest improvement and deterioration for the FLIP evaluator [1] at frames $(1263, 163)$, $(1801, 1107)$, $(1799, 637)$, and $(1899, 1929)$, respectively and are visualized for qualitative inspection in Figure 3.

Figure 4 shows results on the "Room" sequences of [11], they consist of 2642, 1828, and 1789 frames, which include 51, 33, and 34 target frames, respectively, thus incorporating 2591, 1795, and 1755 novel, unseen viewpoints. The "Room 0", "Room 1", and "Room 2" sequences have their largest improvement and deterioration for the $L_2$ error at frames $(1856, 1203)$, $(31, 114)$, and $(604, 118)$, respectively and are visualized for qualitative inspection in Figure 4. The same sequences have their largest improvement and deterioration for the FLIP evaluator [1] at frames $(1552, 1204)$, $(31, 83)$, and $(656, 118)$, respectively and are visualized for qualitative inspection in Figure 5.

Overall, the average reconstruction error decreases for all experiments and validates our findings described in Section 4.2.1 on a larger scale. This can also be seen qualitatively; note the overall increase of realism, for the improvements, due to view-dependent effects, while the deteriorations seem to be only slightly worse than the baseline, but still visually pleasing to the human eye – an effect desired in AR/VR/MR applications.

**Further quantitative results** on the Room sequences of [11] are shown in Figure 6. For specular highlights that seem too wide such as the vase in "Room 0" our reconstructions still look more faithful compared to a purely diffuse one. Notice that the anisotropic BRDF of the window blinds in "Room 2" is difficult to recover with our approach as we do not model this effect. Instead, we estimate an isotropic approximation of it, which still looks realistic.

**Robustness against inaccurate geometry** can affect the final reconstruction in accuracy and realism. Figure 3 "Office 3" shows how specularities are misplaced and BRDF estimates too rough, if the geometry (clock) is inaccurate at the location of reflection. Figure 4 and 5 "Room 1" and Figure 6 "Room 0" and "Room 1" show results were different levels of deteriorated geometry affects the non-diffuse BRDF estimate. While the vase in "Room 1" is almost diffuse, the vase in "Room 0" shows specular reflections, although not as strong as the capture. The reason for both failures are caused by an estimated specular highlight having no overlap with the genuine reflection, cp. the error maps in Figure 6 "Room 0", the specular reflections are not perfectly superimposed.

## E. Further Relighting results (Section 4.3)

Further renderings under novel lighting with artificially placed objects are shown in Figure 7.

## F. Attached video file

The video file attached to the supplementary material shows a number of video renderings of our results as well as comparisons to the baseline. This video is encoded with the H.265 codec in an MP4 container. Some of the images shown in the video will have a somewhat grainy appearance - this is caused by the relatively simple path tracer we implemented for visualizing the results of our approach, rather than being an intrinsic part of the estimated appearance.

## References


[1] P. Andersson, J. Nilsson, T. Akenine-Möller, M. Oskarsson, K. Åström, and M. D. Fairchild. Flip: a difference evaluator for alternating images. *Proceedings of the ACM on Computer Graphics and Interactive Techniques (HPG 2020)*, 3(2), 2020. 2, 5, 7, 8

[2] J. Engel, V. Koltun, and D. Cremers. Direct sparse odometry. *IEEE transactions on pattern analysis and machine intelligence*, 40(3):611–625, 2017. 1

[3] W. Jakob, M. Tarini, D. Panozzo, and O. Sorkine-Hornung. Instant field-aligned meshes. *ACM Trans. Graph.*, 34(6):189–1, 2015. 1


Figure 1. We deploy our albedo and shading estimation on challenging real-world "Room" data sets of the Replica data set [11] and are able to estimate per-texel albedo and shading information, using the reconstructed mesh and lit diffuse HDR texture only.


[4] M. Kazhdan, M. Bolitho, and H. Hoppe. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, volume 7, 2006. 1

[5] M. Kazhdan and H. Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics (ToG)*, 32(3):1–13, 2013. 1

[6] M. McGuire. Computer graphics archive. https://casual-effects.com/data, July 2017. 8

[7] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015. 1

[8] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohli, J. Shotton, S. Hodges, and A. Fitzgibbon. KinectFusion: Real-Time Dense Surface Mapping and Tracking. In *Proceedings of the International Symposium on Mixed and Augmented Reality (ISMAR)*, 2011. 1

[9] M. Pharr, W. Jakob, and G. Humphreys. *Physically based rendering: From theory to implementation*. Morgan Kaufmann, 2016. 1

[10] The Stanford 3D Scanning Repository. http://graphics.stanford.edu/data/3Dscanrep/. 8

[11] J. Straub, T. Whelan, L. Ma, Y. Chen, E. Wijmans, S. Green, J. J. Engel, R. Mur-Artal, C. Ren, S. Verma, A. Clarkson, M. Yan, B. Budge, Y. Yan, X. Pan, J. Yon, Y. Zou, K. Leon, N. Carter, J. Briales, T. Gillingham, E. Mueggler, L. Pesqueira, M. Savva, D. Batra, H. M. Strasdat, R. D. Nardi, M. Goesele, S. Lovegrove, and R. Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 2, 3, 4, 5, 6, 7, 8

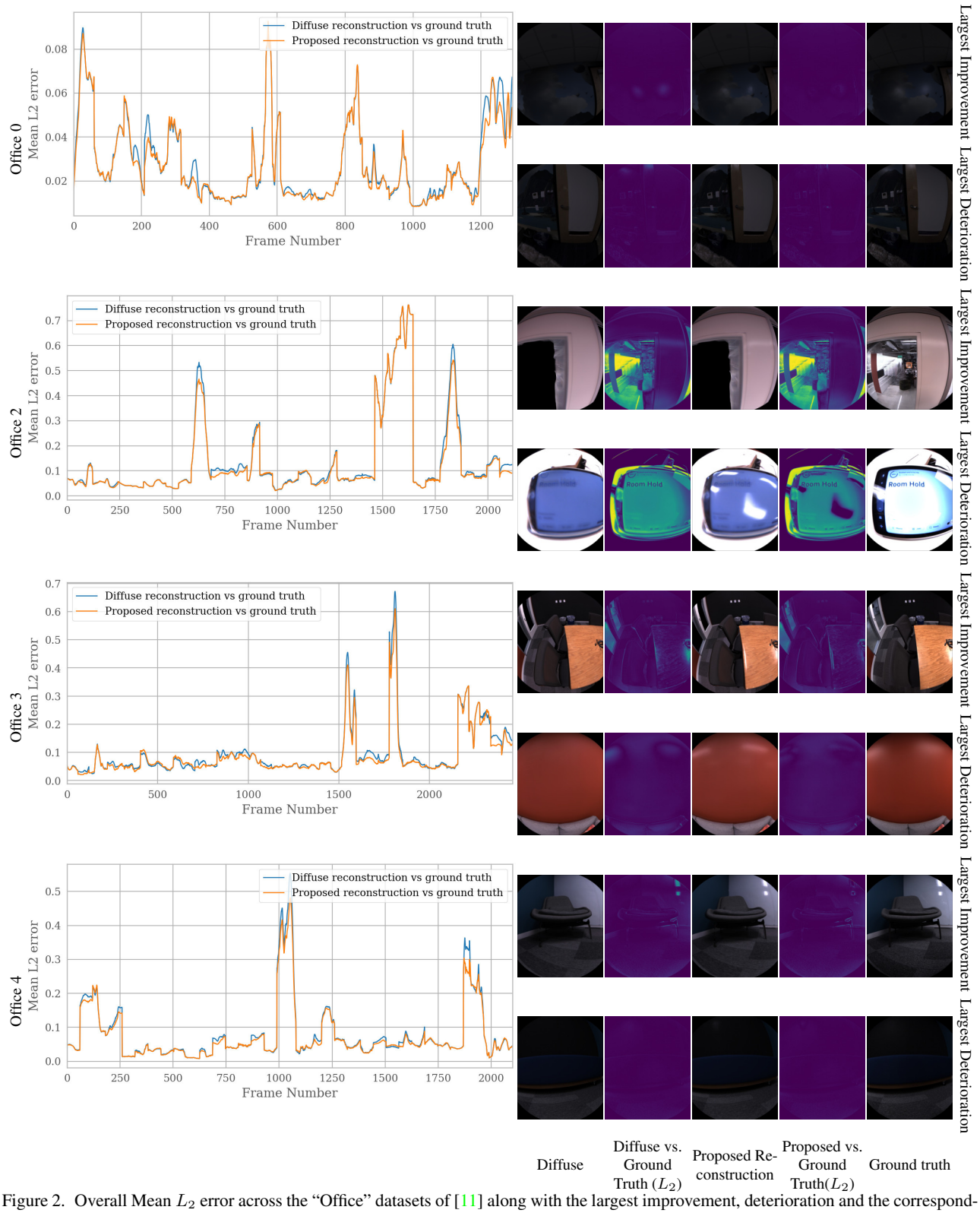



Figure 2. Overall Mean $L_2$ error across the "Office" datasets of [11] along with the largest improvement, deterioration and the corresponding $L_2$ error maps.

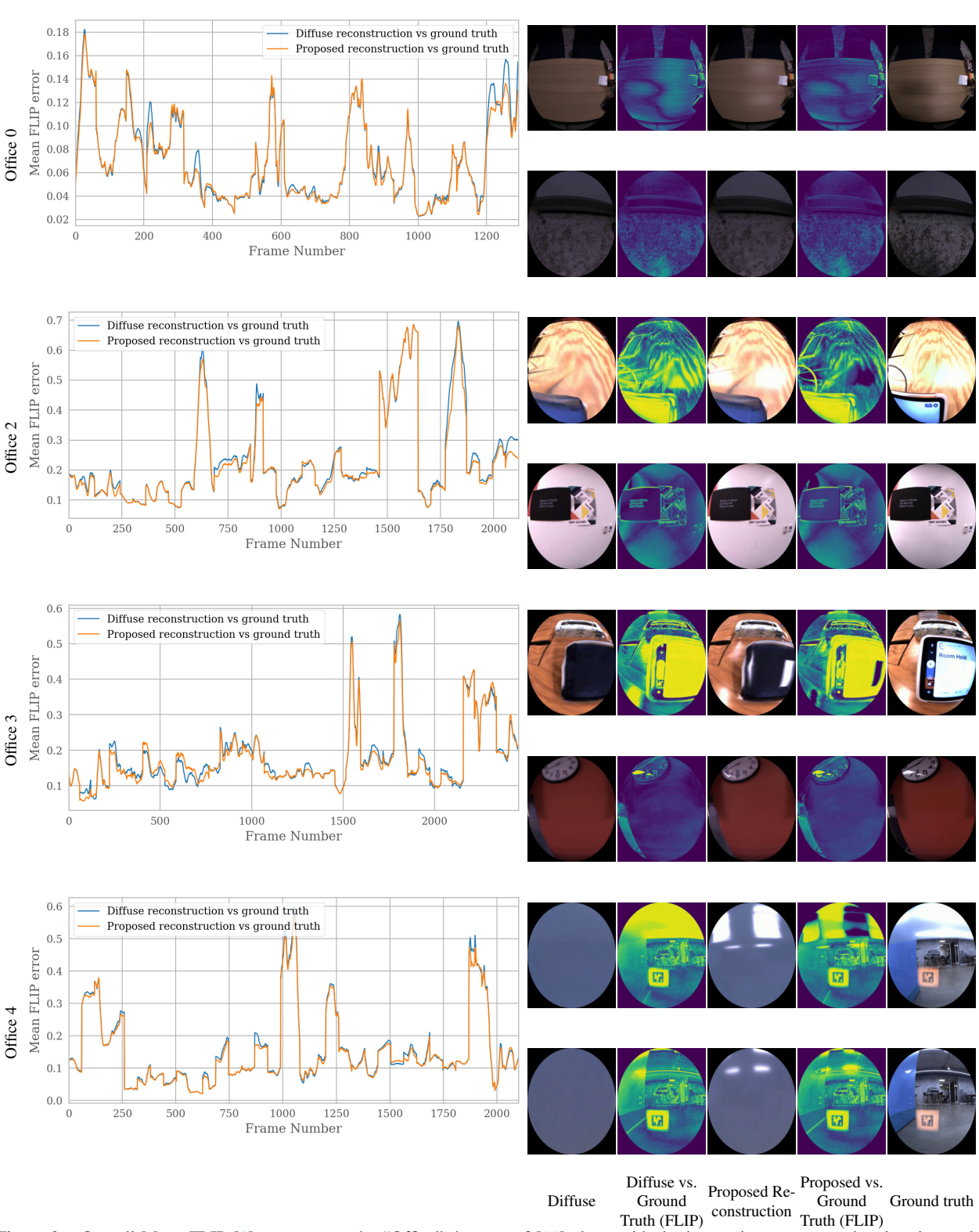



Figure 3. Overall Mean FLIP [1] error across the "Office" datasets of [11] along with the largest improvement, deterioration and the corresponding FLIP error maps.

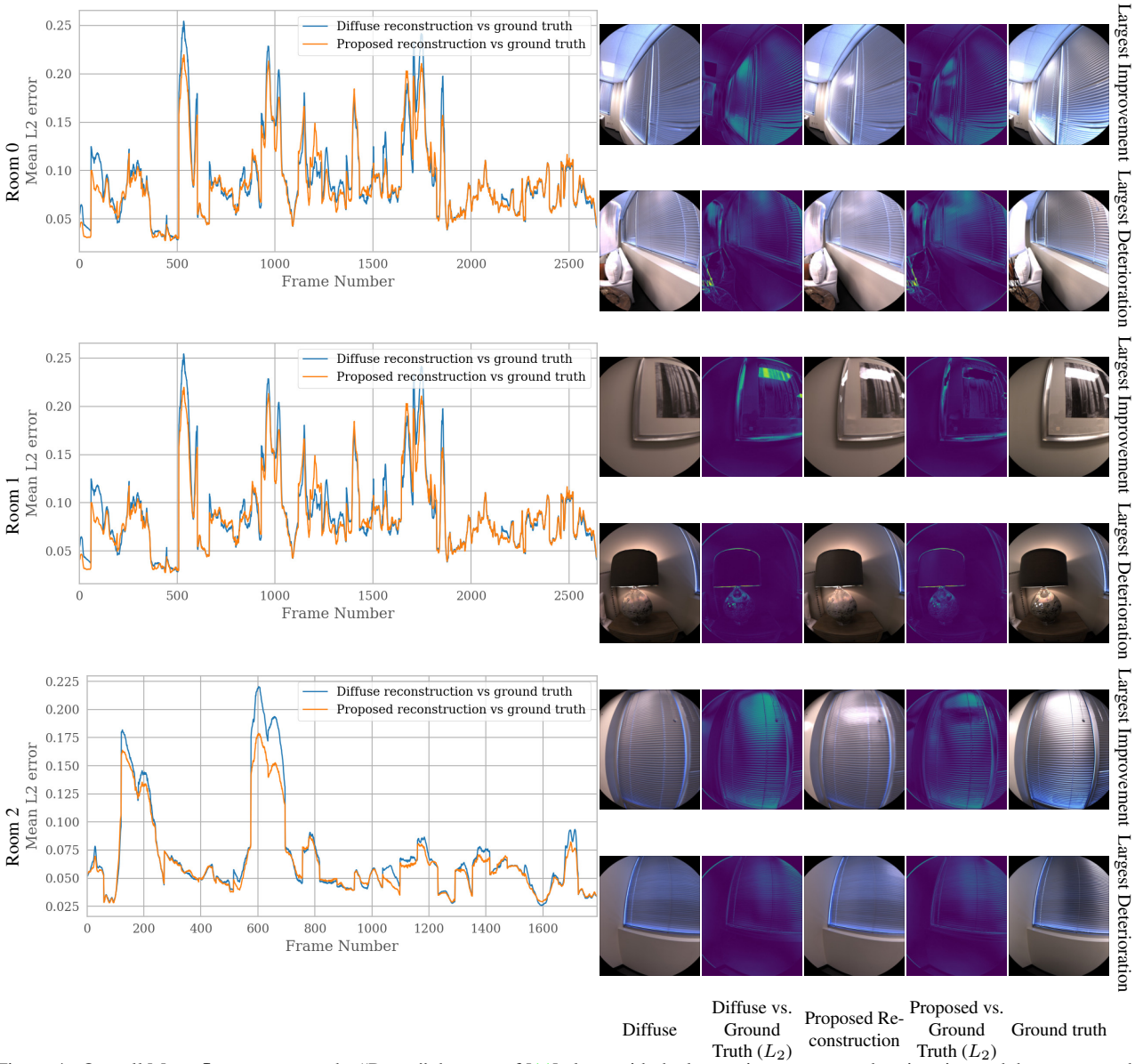



Figure 4. Overall Mean $L_2$ error across the "Room" datasets of [11] along with the largest improvement, deterioration and the corresponding $L_2$ error maps.

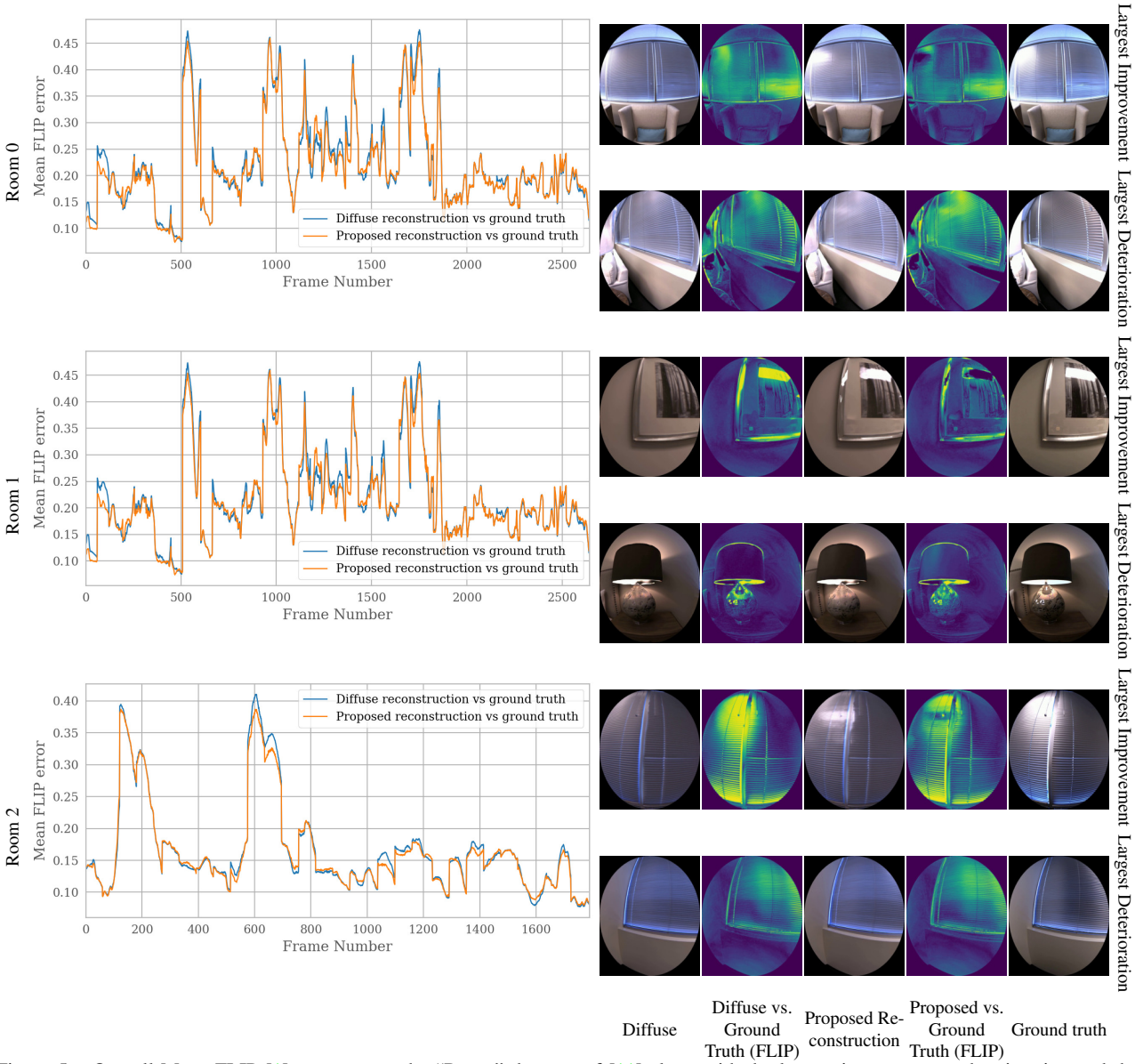



Figure 5. Overall Mean FLIP [1] error across the "Room" datasets of [11] along with the largest improvement, deterioration and the corresponding FLIP error maps.

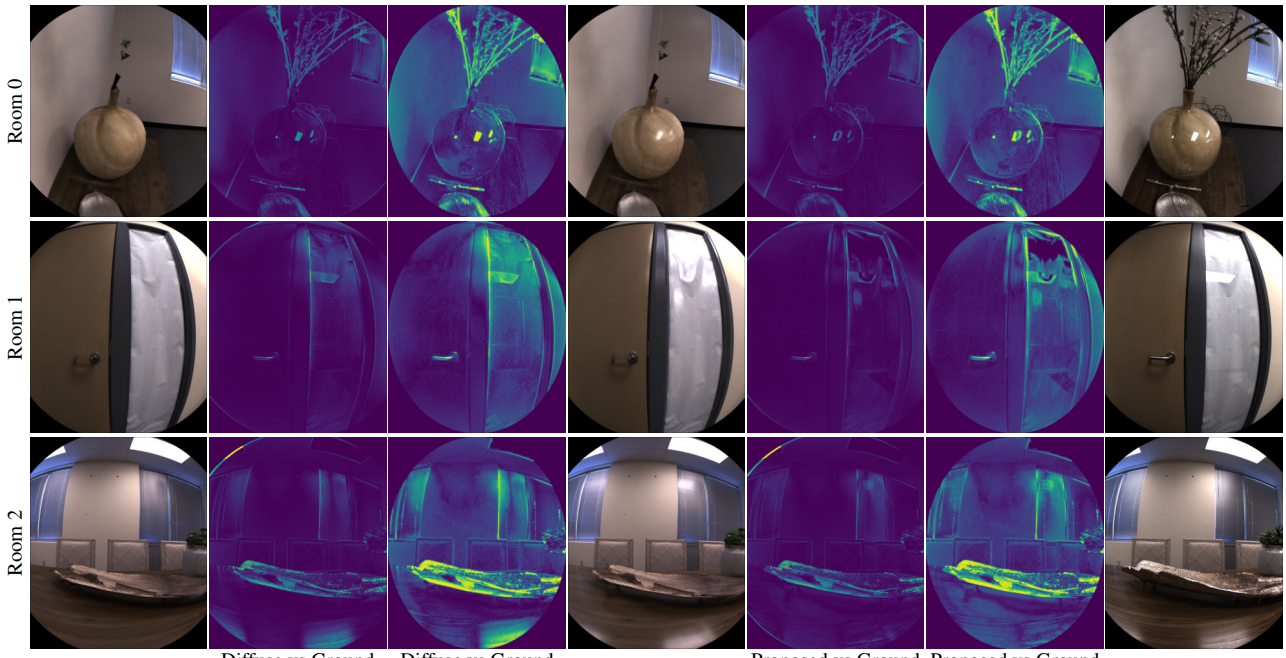



Figure 6. Side-by-side comparisons between the diffuse and the proposed reconstruction along with the ground truth and the corresponding $L_2$ errors and FLIP evaluator [1]. Adding the proposed specular appearance estimate makes reconstructions more realistic.

Figure 7. Complete synthetic relighting of different data sets (Office 0, Room 0 and Room 1 [11]) with additional virtually placed objects [10, 6].

# Licenses

## Fight Ill-Posedness with Ill-Posedness: Single-shot Variational Depth Super-Resolution from Shading

**Conference Proceedings:**
2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition

**Author:** Bjoern Haefner

**Publisher:** IEEE

**Date:** June 2018



### Thesis / Dissertation Reuse

**The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:**

*Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:*

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

*Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:*

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

BACK CLOSE WINDOW

## Variational Uncalibrated Photometric Stereo Under General Lighting

**Conference Proceedings:** 2019 IEEE/CVF International Conference on Computer Vision (ICCV)

**Author:** Bjoern Haefner

**Publisher:** IEEE

**Date:** October 2019



### Thesis / Dissertation Reuse

**The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:**

*Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:*

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

*Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:*

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

BACK CLOSE WINDOW

IEEE Requesting permission to reuse content from an IEEE publication

## Photometric Segmentation: Simultaneous Photometric Stereo and Masking

**Conference Proceedings:** 2019 International Conference on 3D Vision (3DV)

**Author:** Bjoern Haefner

**Publisher:** IEEE

**Date:** September 2019



### Thesis / Dissertation Reuse

**The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:**

*Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:*

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

*Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:*

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

IEEE Requesting permission to reuse content from an IEEE publication

## Recovering Real-World Reflectance Properties and Shading From HDR Imagery

**Conference Proceedings:** 2021 International Conference on 3D Vision (3DV)

**Author:** Bjoern Haefner

**Publisher:** IEEE

**Date:** December 2021



### Thesis / Dissertation Reuse

**The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:**

*Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:*

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

*Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:*

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

# List of Figures

# List of Tables

# Acronyms

**AR** Augmented Reality. 3, 5

**BRDF** Bidirectional Reflectance Distribution Function. 5–7, 17, 18, 20–25, 27, 28, 70–73, 131, 132, 137–140

**BSDF** Bidirectional Scattering Distribution Function. 140

**BTDF** Bidirectional Transmittance Distribution Function. 22, 137, 140

**CPS** Calibrated Photometric Stereo. 48–50, 131, 139

**dof** Degree of Freedom. 49

**EXIF** Exchangeable Image File Format. 11

**FPS** Frames per Second. 135

**GBR** Generalized Bas-Relief. 49

**GPU** Graphics Processing Unit. 20, 62, 136, 138

**HDR** High Dynamic Range. 77–79, 132

**HSV** Hue, Saturation, Value. 138

**IMU** Inertial Measurement Unit. 135

**IOR** Index of Refraction. 26

**IR** Infrared. 12, 138

**LR** Low-Resolution. 4, 14, 15, 45, 73, 77, 131

**MLP** Multilayer Perceptron. 137

**MR** Mixed Reality. 3, 5

**MRF** Markov Random Field. 58

**MVS** Multi View Stereo. 66, 67

**NLM** Non-Local Mean. 58

**PDE** Partial Differential Equation. 45, 50, 51

**PMD** Photonic Mixer Device. 12

**PS** Photometric Stereo. 4–7, 9, 17, 21, 41, 42, 46–50, 52, 53, 55, 57, 59, 66, 69, 70, 77, 78, 131, 132, 136, 137, 139

**SDF** Signed Distance Function. 37, 138

**SfM** Structure from Motion. 66, 67

**SfS** Shape-from-Shading. 4–7, 9, 17, 21, 41–47, 57, 59–61, 65, 69, 77, 131

**SGD** Stochastic Gradient Descent. 71, 136

**SH** Spherical Harmonics. 31–37, 41, 48, 49, 52, 60–67

**SR** Super-Resolution. 5–7, 13–17, 57–59, 65, 67, 77, 131, 135, 138

**SVBRDF** Spatially Varying Bidirectional Reflectance Distribution Function. 22, 139

**SVD** Singular Value Decomposition. 65–68

**TGV** Total Generalized Variation. 59

**TOF** Time-of-Flight. 11–14

**TV** Total Variation. 16, 59, 78

**UPS** Uncalibrated Photometric Stereo. 6, 48, 49, 52, 57, 65–68, 78, 131, 132, 138

**VR** Virtual Reality. 3, 5

# Own Publications


[1] M. Brahimi, Y. Quéau, B. Haefner, and D. Cremers. *On the Well-Posedness of Uncalibrated Photometric Stereo Under General Lighting*. In *Advances in Photometric 3D-Reconstruction*. J.-D. Durou, M. Falcone, Y. Quéau, and S. Tozza, editors. Springer International Publishing, 2020, pages 147–176 (cited on pp. 39, 40, 46, 47, 71–73).

[2] B. Haefner, S. Green, A. Oursland, D. Andersen, M. Goesele, D. Cremers, R. Newcombe, and T. Whelan. Recovering Real-world Reflectance Properties and Shading from HDR Imagery. In *International Conference on 3D Vision (3DV)*, 2021 (cited on pp. 8, 20, 69, 71–73, 111).

[3] B. Haefner, S. Peng, A. Verma, Y. Quéau, and D. Cremers. Photometric Depth Super-Resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 42(10):2453–2464, 2020 (cited on pp. 16, 33, 39, 71, 72).

[4] B. Haefner, Y. Quéau, and D. Cremers. Photometric Segmentation: Simultaneous Photometric Stereo and Masking. In *International Conference on 3D Vision (3DV)*, 2019 (cited on pp. 6, 8, 39, 71–73, 101).

[5] B. Haefner, Y. Quéau, T. Möllenhoff, and D. Cremers. Fight ill-posedness with ill-posedness: Single-shot variational depth super-resolution from shading. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018 (cited on pp. 6, 8, 16, 33, 39, 71–73, 77).

[6] B. Haefner, Z. Ye, M. Gao, T. Wu, Y. Quéau, and D. Cremers. Variational Uncalibrated Photometric Stereo under General Lighting. In *International Conference on Computer Vision (ICCV)*, 2019 (cited on pp. 6, 8, 33, 34, 39, 40, 71–73, 89).

[7] S. Peng, B. Haefner, Y. Quéau, and D. Cremers. Depth Super-Resolution Meets Uncalibrated Photometric Stereo. In *International Conference on Computer Vision (ICCV) Workshops*, 2017 (cited on pp. 15, 16, 33, 39, 50, 57, 63–65, 71, 72, 74).

[8] L. Sang, B. Haefner, and D. Cremers. Inferring Super-Resolution Depth from a Moving Light-Source Enhanced RGB-D Sensor: A Variational Approach. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020 (cited on pp. 16, 33, 39, 45, 57, 71, 72).

[9] L. Sang, B. Haefner, X. Zuo, and D. Cremers. High-Quality RGB-D Reconstruction via Multi-View Uncalibrated Photometric Stereo and Gradient-SDF. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2023 (cited on pp. 33, 45, 71, 72).

[10] Z. Ye, B. Haefner, Y. Quéau, T. Möllenhoff, and D. Cremers. A Cutting-Plane Method for Sublabel-Accurate Relaxation of Problems with Product Label Spaces. *International Journal of Computer Vision (IJCV)*, 2022 (cited on pp. 71, 72, 128).

[11] Z. Ye, B. Haefner, Y. Quéau, T. Möllenhoff, and D. Cremers. Sublabel-Accurate Multilabeling Meets Product Label Spaces. In *German Conference on Pattern Recognition (GCPR)*, 2021 (cited on pp. 71, 72, 128).

# Bibliography


[12] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. Slic superpixels. Technical report, 2010 (cited on p. 56).

[13] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012 (cited on p. 56).

[14] E. Adelson and A. Pentland. *The perception of shading and reflectance*. In *Perception as Bayesian Inference*. D. C. Knill and W. Richards, editors. Cambridge University Press, 1996, pages 409–424. DOI: 10.1017/CBO9780511984037.014 (cited on pp. 40–42).

[15] T. Akenine-Möller, E. Haines, N. Hoffman, A. Pesce, M. Iwanicki, and S. Hillaire. *Real-Time Rendering 4th Edition*. A K Peters/CRC Press, Boca Raton, FL, USA, 2018, page 1200. ISBN: 978-1-13862-700-0 (cited on pp. 17, 20, 23).

[16] E. Alexander, D. Holtmann-Rice, and S. Zucker. When shading flows with color: grounding shape and material inference. *Journal of Vision*, 13(9):257–257, 2013 (cited on p. 129).

[17] T. D. Alter. 3d pose from three corresponding points under weak-perspective projection, 1992 (cited on p. 8).

[18] D. Azinovic, T.-M. Li, A. Kaplanyan, and M. Nießner. Inverse Path Tracing for Joint Material and Lighting Estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2442–2451, 2019 (cited on pp. 68–70, 74, 128).

[19] J. T. Barron and J. Malik. Shape, illumination, and reflectance from shading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(8):1670–1687, 2015. DOI: 10.1109/TPAMI.2014.2377712 (cited on p. 4).

[20] R. Basri, D. Jacobs, and I. Kemelmacher. Photometric stereo with general, unknown lighting. *International Journal of computer vision*, 72:239–257, 2007 (cited on pp. 4, 40, 47, 50, 63–65).

[21] R. Basri and D. W. Jacobs. Lambertian reflectance and linear subspaces. *IEEE transactions on pattern analysis and machine intelligence*, 25(2):218–233, 2003 (cited on pp. 30, 32, 33, 49).

[22] J. Batlle, E. Mouaddib, and J. Salvi. Recent progress in coded structured light as a technique to solve the correspondence problem: a survey. *Pattern recognition*, 31(7):963–982, 1998 (cited on p. 11).

[23] P. N. Belhumeur, D. J. Kriegman, and A. L. Yuille. The bas-relief ambiguity. *International journal of computer vision*, 35(1):33–44, 1999 (cited on pp. 46, 65).

[24] O. Ben-Shahar and S. W. Zucker. Hue geometry and horizontal connections. *Neural Networks*, 17(5-6):753–771, 2004 (cited on p. 129).

[25] P. Bergmann, R. Wang, and D. Cremers. Online photometric calibration of auto exposure video for realtime visual odometry and slam. *IEEE Robotics and Automation Letters*, 3(2):627–634, 2017 (cited on p. 19).

[26] P. Bhat, C. L. Zitnick, M. Cohen, and B. Curless. Gradientshop: a gradient-domain optimization framework for image and video filtering. *ACM Transactions on Graphics (TOG)*, 29(2):1–14, 2010 (cited on p. 56).

[27] S. Bi, Z. Xu, K. Sunkavalli, M. Hašan, Y. Hold-Geoffroy, D. Kriegman, and R. Ramamoorthi. Deep reflectance volumes: relightable reconstructions from multi-view photometric images. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 294–311. Springer, 2020 (cited on p. 45).

[28] M. A. Blanco, M. Flórez, and M. Bermejo. Evaluation of the rotation matrices in the basis of real spherical harmonics. *Journal of Molecular Structure: THEOCHEM*, 419(1-3):19–27, 1997 (cited on pp. 30–32).

[29] J. F. Blinn. Models of light reflection for computer synthesized pictures. In *Proceedings of the 4th annual conference on Computer graphics and interactive techniques*, pages 192–198, 1977 (cited on pp. 21, 22).

[30] L. Bode, S. Merzbach, P. Stotko, M. Weinmann, and R. Klein. Real-time multi-material reflectance reconstruction for large-scale scenes under uncontrolled illumination from rgb-d image sequences. In *2019 International Conference on 3D Vision (3DV)*, pages 709–718. IEEE, 2019 (cited on p. 20).

[31] M. Boss, A. Engelhardt, A. Kar, Y. Li, D. Sun, J. Barron, H. Lensch, and V. Jampani. Samurai: shape and material from unconstrained real-world arbitrary image collections. *Advances in Neural Information Processing Systems*, 35:26389–26403, 2022 (cited on pp. 40, 128).

[32] M. Boss, V. Jampani, K. Kim, H. Lensch, and J. Kautz. Two-shot spatially-varying brdf and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3982–3991, 2020 (cited on p. 67).

[33] K. Bredies, K. Kunisch, and T. Pock. Total generalized variation. *SIAM Journal on Imaging Sciences*, 3(3):492–526, 2010 (cited on p. 57).

[34] M. Breuß, E. Cristiani, J.-D. Durou, M. Falcone, and O. Vogel. Perspective shape from shading: ambiguity analysis and numerical approximations. *SIAM Journal on Imaging Sciences*, 5(1):311–342, 2012 (cited on p. 43).

[35] A. R. Bruss. The eikonal equation: some results applicable to computer vision. *Journal of Mathematical Physics*, 23(5):890–896, 1982 (cited on p. 42).

[36] A. Buades, B. Coll, and J.-M. Morel. A non-local algorithm for image denoising. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 2, pages 60–65. Ieee, 2005 (cited on p. 56).

[37] B. Burley. Extending the disney brdf to a bsdf with integrated subsurface scattering. *Physically Based Shading in Theory and Practice'SIGGRAPH Course*, 2015 (cited on p. 131).

[38] B. Burley and W. D. A. Studios. Physically-based shading at disney. In *Acm Siggraph*, volume 2012, pages 1–7. vol. 2012, 2012 (cited on pp. 3, 21, 22, 24, 26, 27, 129).

[39] J. Cao, H. Wang, P. Chemerys, V. Shakhrai, J. Hu, Y. Fu, D. Makoviichuk, S. Tulyakov, and J. Ren. Real-time neural light field on mobile devices, 2023 (cited on p. 3).

[40] K. C. Chan, S. Zhou, X. Xu, and C. C. Loy. Basicvsr++: improving video superresolution with enhanced propagation and alignment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5972–5981, 2022 (cited on p. 14).

[41] T. F. Chan and L. A. Vese. Active contours without edges. *IEEE Transactions on Image Processing (TIP)*, 10(2):266–277, 2001 (cited on pp. 50, 51, 67).

[42] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang. Matterport3d: learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017 (cited on p. 69).

[43] 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*, 44(1):129–142, 2020 (cited on pp. 45, 128).

[44] G. Chen, K. Han, B. Shi, Y. Matsushita, and K.-Y. K. Wong. Self-calibrating deep photometric stereo networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8739–8747, 2019 (cited on p. 128).

[45] G. Chen, M. Waechter, B. Shi, K.-Y. K. Wong, and Y. Matsushita. What is learned in deep uncalibrated photometric stereo? In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 745–762. Springer, 2020 (cited on p. 128).

[46] J. Chen, C.-K. Tang, and J. Wang. Noise brush: interactive high quality image-noise separation. In *ACM SIGGRAPH Asia 2009 papers*, pages 1–10. 2009 (cited on p. 56).

[47] L. Chen, Y. Zheng, B. Shi, A. Subpa-Asa, and I. Sato. A microfacet-based model for photometric stereo with general isotropic reflectance. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):48–61, 2019 (cited on p. 67).

[48] K. H. Cheng and A. Kumar. Revisiting outlier rejection approach for non-lambertian photometric stereo. *IEEE Transactions on Image Processing*, 28(3):1544–1555, 2018 (cited on p. 45).

[49] C. D. H. Chisholm. Group theoretical techniques in quantum chemistry. 1976 (cited on p. 30).

[50] T. Collins and A. Bartoli. 3d reconstruction in laparoscopy with close-range photometric stereo. In *MICCAI (2)*, pages 634–642. Citeseer, 2012 (cited on p. 3).

[51] R. L. Cook and K. E. Torrance. A reflectance model for computer graphics. *ACM Transactions on Graphics (ToG)*, 1(1):7–24, 1982 (cited on pp. 22, 24, 129).

[52] E. Cristiani and M. Falcone. Fast semi-lagrangian schemes for the eikonal equation and applications. *SIAM Journal on Numerical Analysis*, 45(5):1979–2011, 2007 (cited on p. 43).

[53] J. G. da Silva Neto, P. J. da Lima Silva, F. Figueredo, J. M. X. N. Teixeira, and V. Teichrieb. Comparison of rgb-d sensors for 3d reconstruction. In *2020 22nd Symposium on Virtual and Augmented Reality (SVR)*, pages 252–261. IEEE, 2020 (cited on p. 13).

[54] P. A. Davis and L. A. Soderblom. Modeling crater topography and albedo from monoscopic viking orbiter images: 1. methodology. *Journal of Geophysical Research: Solid Earth*, 89(B11):9449–9457, 1984 (cited on p. 48).

[55] P. E. Debevec and J. Malik. Recovering high dynamic range radiance maps from photographs. In *ACM SIGGRAPH 2008 classes*, pages 1–10. 2008 (cited on p. 19).

[56] V. Deschaintre, M. Aittala, F. Durand, G. Drettakis, and A. Bousseau. Single-image svbrdf capture with a rendering-aware deep network. *ACM Transactions on Graphics (ToG)*, 37(4):1–15, 2018 (cited on p. 67).

[57] F. Devernay and O. Faugeras. Straight lines have to be straight. *Machine vision and applications*, 13:14–24, 2001 (cited on p. 8).

[58] J. Diebel and S. Thrun. An application of markov random fields to range sensing. *Advances in neural information processing systems*, 18, 2005 (cited on pp. 16, 55–57).

[59] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015 (cited on p. 14).

[60] J.-D. Durou, M. Falcone, and M. Sagona. Numerical methods for shape-from-shading: a new survey with benchmarks. *Computer Vision and Image Understanding*, 109(1):22–43, 2008 (cited on p. 43).

[61] A. Einstein. Zur elektrodynamik bewegter körper. *Annalen der physik*, 4, 1905 (cited on p. 47).

[62] R. Or-El, R. Hershkovitz, A. Wetzler, G. Rosman, A. M. Bruckstein, and R. Kimmel. Real-time depth refinement for specular objects. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4378–4386, 2016 (cited on p. 130).

[63] M. Elad and A. Feuer. Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images. *IEEE transactions on image processing*, 6(12):1646–1658, 1997 (cited on p. 14).

[64] J. Engel, V. Usenko, and D. Cremers. A photometrically calibrated benchmark for monocular visual odometry. *arXiv preprint arXiv:1607.02555*, 2016 (cited on p. 19).

[65] C. H. Esteban, G. Vogiatzis, and R. Cipolla. Multiview photometric stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(3):548–554, 2008 (cited on p. 44).

[66] M. Falcone and M. Sagona. An algorithm for the global solution of the shape-from-shading model. In *Image Analysis and Processing: 9th International Conference, ICIAP'97 Florence, Italy, September 17–19, 1997 Proceedings, Volume I 9*, pages 596–603. Springer, 1997 (cited on p. 43).

[67] A. R. Farooq, M. L. Smith, L. N. Smith, and S. Midha. Dynamic photometric stereo for on line quality control of ceramic tiles. *Computers in industry*, 56(8-9):918–934, 2005 (cited on p. 3).

[68] P. Favaro and T. Papadhimitri. A closed-form solution to uncalibrated photometric stereo via diffuse maxima. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 821–828, 2012 (cited on p. 74).

[69] R. Fernando. *GPU Gems: Programming Techniques, Tips and Tricks for Real-Time Graphics*. Pearson Higher Education, 2004. ISBN: 0321228324 (cited on p. 20).

[70] D. Ferstl, C. Reinbacher, R. Ranftl, M. Rüther, and H. Bischof. Image guided depth upsampling using anisotropic total generalized variation. In *Proceedings of the IEEE international conference on computer vision*, pages 993–1000, 2013 (cited on pp. 16, 55–57).

[71] R. T. Frankot and R. Chellappa. A method for enforcing integrability in shape from shading algorithms. *IEEE Transactions on pattern analysis and machine intelligence*, 10(4):439–451, 1988 (cited on p. 43).

[72] D. Frolova, D. Simakov, and R. Basri. Accuracy of spherical harmonic approximations for images of lambertian objects under far and near lighting. In *Computer Vision-ECCV 2004: 8th European Conference on Computer Vision, Prague, Czech Republic, May 11-14, 2004. Proceedings, Part I 8*, pages 574–587. Springer, 2004 (cited on pp. 30, 35).

[73] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1362–1376, 2009 (cited on p. 64).

[74] D. Gao, X. Li, Y. Dong, P. Peers, K. Xu, and X. Tong. Deep inverse rendering for high-resolution svbrdf estimation from an arbitrary number of images. *ACM Trans. Graph.*, 38(4):134–1, 2019 (cited on pp. 5, 67).

[75] J. Geng. Structured-light 3d surface imaging: a tutorial. *Advances in Optics and Photonics*, 3(2):128–160, 2011 (cited on p. 12).

[76] S. Giancola, M. Valenti, and R. Sala. *A survey on 3D cameras: Metrological comparison of time-of-flight, structured-light and active stereoscopy technologies*. Springer, 2018 (cited on pp. 11–13).

[77] B. Goldluecke, E. Strekalovskiy, and D. Cremers. Tight convex relaxations for vector-valued labeling. *SIAM Journal on Imaging Sciences*, 6(3):1626–1664, 2013 (cited on p. 128).

[78] P. F. Gotardo, T. Simon, Y. Sheikh, and I. Matthews. Photogeometric scene flow for high-detail dynamic 3d reconstruction. In *Proceedings of the IEEE international conference on computer vision*, pages 846–854, 2015 (cited on pp. 39, 45, 48).

[79] G. Graber, J. Balzer, S. Soatto, and T. Pock. Efficient minimal-surface regularization of perspective depth maps in variational stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 511–520, 2015 (cited on pp. 16, 38, 62).

[80] H. Groemer. *Geometric applications of Fourier series and spherical harmonics*, volume 61. Cambridge University Press, 1996 (cited on p. 32).

[81] R. Grosse, M. K. Johnson, E. H. Adelson, and W. T. Freeman. Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *2009 IEEE 12th International Conference on Computer Vision*, pages 2335–2342. IEEE, 2009 (cited on p. 60).

[82] G. Guidi, S. GONIZZI BARSANTI, L. L. Micoli, et al. 3d capturing performances of low-cost range sensors for mass-market applications. *International archives of the photogrammetry, remote sensing and spatial information sciences*:33–40, 2016 (cited on p. 13).

[83] J. Gunther, S. Popov, H.-P. Seidel, and P. Slusallek. Realtime ray tracing on gpu with bvh-based packet traversal. In *2007 IEEE Symposium on Interactive Ray Tracing*, pages 113–118, 2007. DOI: 10.1109/RT.2007.4342598 (cited on p. 20).

[84] H. Guo, H. Santo, B. Shi, and Y. Matsushita. Edge-preserving near-light photometric stereo with neural surfaces. *arXiv preprint arXiv:2207.04622*, 2022 (cited on p. 128).

[85] T. Hach and J. Steurer. A novel rgb-z camera for high-quality motion picture applications. In *Proceedings of the 10th European Conference on Visual Media Production*, pages 1–10, 2013 (cited on p. 127).

[86] Y. Han, J.-Y. Lee, and I. So Kweon. High quality shape from a single rgb-d image under uncalibrated natural illumination. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1617–1624, 2013 (cited on pp. 39, 43, 57–62).

[87] C. Hardy, Y. Quéau, and D. Tschumperlé. Ms-ps: a multi-scale network for photometric stereo with a new comprehensive training dataset. *arXiv preprint arXiv:2211.14118*, 2022 (cited on p. 128).

[88] H. Hayakawa. Photometric stereo under a light source with arbitrary motion. *JOSA A*, 11(11):3079–3089, 1994 (cited on pp. 40, 46, 65).

[89] E. Heitz. Understanding the masking-shadowing function in microfacet-based brdfs. *Journal of Computer Graphics Techniques*, 3(2):32–91, 2014 (cited on p. 27).

[90] C. Hernández, G. Vogiatzis, G. J. Brostow, B. Stenger, and R. Cipolla. Non-rigid photometric stereo with colored lights. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007 (cited on p. 45).

[91] D. Holtmann-Rice, E. Alexander, R. Fleming, and S. Zucker. When color flows with shading: making depth disappear. *Journal of Vision*, 13(9):467–467, 2013 (cited on p. 129).

[92] B. K. P. Horn. *Shape From Shading: A Method for Obtaining the Shape of a Smooth Opaque Object From One View*. PhD thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 1970 (cited on pp. 4, 41, 66).

[93] B. Horn, B. Klaus, and P. Horn. *Robot vision*. MIT press, 1986 (cited on pp. 18, 19).

[94] B. K. Horn and M. J. Brooks. The variational approach to shape from shading. *Computer Vision, Graphics, and Image Processing*, 33(2):174–208, 1986 (cited on pp. 43, 59).

[95] R. Huang and W. A. Smith. Shape-from-shading under complex natural illumination. In *2011 18th IEEE International Conference on Image Processing*, pages 13–16. IEEE, 2011 (cited on p. 43).

[96] P. J. Huber. Robust estimation of a location parameter. *Breakthroughs in statistics: Methodology and distribution*:492–518, 1992 (cited on p. 16).

[97] S. Ikehata. Scalable, detailed and mask-free universal photometric stereo. *arXiv preprint arXiv:2303.15724*, 2023 (cited on pp. 128, 131).

[98] S. Ikehata. Universal photometric stereo network using global lighting contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12591–12600, 2022 (cited on p. 128).

[99] D. S. Immel, M. F. Cohen, and D. P. Greenberg. A radiosity method for non-diffuse environments. *Acm Siggraph Computer Graphics*, 20(4):133–142, 1986 (cited on p. 17).

[100] Immel, David S. *A radiosity method for non-diffuse surfaces*. Master's thesis, Cornell University, 1986 (cited on p. 17).

[101] Z. Jankó, A. Delaunoy, and E. Prados. Colour dynamic photometric stereo for textured surfaces. In *Computer Vision–ACCV 2010: 10th Asian Conference on Computer Vision, Queenstown, New Zealand, November 8-12, 2010, Revised Selected Papers, Part II 10*, pages 55–66. Springer, 2011 (cited on p. 45).

[102] M. K. Johnson and E. H. Adelson. Shape estimation in natural illumination. In *CVPR 2011*, pages 2553–2560. IEEE, 2011 (cited on p. 43).

[103] Y. Ju, B. Shi, M. Jian, L. Qi, J. Dong, and K.-M. Lam. Normattention-psn: a high-frequency region enhanced photometric stereo network with normalized attention. *International Journal of Computer Vision*, 130(12):3014–3034, 2022 (cited on p. 128).

[104] Y. C. Ju, A. Bruhn, and M. Breuß. Variational perspective shape from shading. In *Scale Space and Variational Methods in Computer Vision: 5th International Conference, SSVM 2015, Lège-Cap Ferret, France, May 31-June 4, 2015, Proceedings 5*, pages 538–550. Springer, 2015 (cited on p. 43).

[105] J. Jung, J.-Y. Lee, and I. S. Kweon. One-day outdoor photometric stereo using skylight estimation. *International Journal of Computer Vision*, 127:1126–1142, 2019 (cited on p. 30).

[106] J. T. Kajiya. The rendering equation. In *Proceedings of the 13th annual conference on Computer graphics and interactive techniques*, pages 143–150, 1986 (cited on pp. 17, 20).

[107] K. Kang, Z. Chen, J. Wang, K. Zhou, and H. Wu. Efficient reflectance capture using an autoencoder. *ACM Trans. Graph.*, 37(4):127–1, 2018 (cited on p. 67).

[108] J. Kannala and S. S. Brandt. A generic camera model and calibration method for conventional, wide-angle, and fish-eye lenses. *IEEE transactions on pattern analysis and machine intelligence*, 28(8):1335–1340, 2006 (cited on p. 8).

[109] H. C. Karaimer and M. S. Brown. A software platform for manipulating the camera imaging pipeline. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 429–444. Springer, 2016 (cited on p. 19).

[110] B. Karis and E. Games. Real shading in unreal engine 4. *Proc. Physically Based Shading Theory Practice*, 4(3):1, 2013 (cited on p. 3).

[111] K. Karsch, V. Hedau, D. Forsyth, and D. Hoiem. Rendering synthetic objects into legacy photographs. *ACM Transactions on Graphics (TOG)*, 30(6):1–12, 2011 (cited on p. 67).

[112] K. Karsch, K. Sunkavalli, S. Hadap, N. Carr, H. Jin, R. Fonte, and M. Sittig. Automatic scene inference for 3d object compositing. *arXiv preprint arXiv:1912.12297*, 2019 (cited on p. 67).

[113] B. Kaya, S. Kumar, C. Oliveira, V. Ferrari, and L. Van Gool. Uncertainty-aware deep multi-view photometric stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12601–12611, 2022 (cited on p. 44).

[114] B. Kaya, S. Kumar, F. Sarno, V. Ferrari, and L. Van Gool. Neural radiance fields approach to deep multi-view photometric stereo. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1965–1977, 2022 (cited on p. 45).

[115] M. Kazhdan, M. Bolitho, and H. Hoppe. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, volume 7, page 0, 2006 (cited on p. 64).

[116] B. Khomutenko, G. Garcia, and P. Martinet. An enhanced unified camera model. *IEEE Robotics and Automation Letters*, 1(1):137–144, 2015 (cited on p. 8).

[117] K. Khoshelham and S. O. Elberink. Accuracy and resolution of kinect depth data for indoor mapping applications. *sensors*, 12(2):1437–1454, 2012 (cited on p. 12).

[118] H. Kim, B. Wilburn, and M. Ben-Ezra. Photometric stereo for dynamic surface orientations. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part I 11*, pages 59–72. Springer, 2010 (cited on p. 45).

[119] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2015 (cited on pp. 68–70).

[120] A. Kolb, E. Barth, R. Koch, and R. Larsen. Time-of-flight cameras in computer graphics. In *Computer Graphics Forum*, volume 29 of number 1, pages 141–159. Wiley Online Library, 2010 (cited on p. 12).

[121] V. Kuleshov, S. Z. Enam, and S. Ermon. Audio super-resolution using neural nets. In *ICLR (Workshop Track)*, 2017 (cited on p. 14).

[122] B. Kunsberg, D. Holtmann-Rice, E. Alexander, S. Cholewiak, R. Fleming, and S. W. Zucker. Colour, contours, shading and shape: flow interactions reveal anchor neighbourhoods. *Interface focus*, 8(4):20180019, 2018 (cited on p. 129).

[123] G. Kurillo, E. Hemingway, M.-L. Cheng, and L. Cheng. Evaluating the accuracy of the azure kinect and kinect v2. *Sensors*, 22(7):2469, 2022 (cited on p. 13).

[124] A. Levy, M. Matthews, M. Sela, G. Wetzstein, and D. Lagun. Melon: nerf with unposed images using equivalence class estimation. *arXiv preprint arXiv:2303.08096*, 2023 (cited on p. 128).

[125] M. Li, Z. Zhou, Z. Wu, B. Shi, C. Diao, and P. Tan. Multi-view photometric stereo: a robust solution and benchmark dataset for spatially varying isotropic materials. *IEEE Transactions on Image Processing*, 29:4159–4173, 2020 (cited on p. 45).

[126] T.-M. Li, M. Aittala, F. Durand, and J. Lehtinen. Differentiable monte carlo ray tracing through edge sampling. *ACM Transactions on Graphics (TOG)*, 37(6):1–11, 2018 (cited on pp. 40, 128).

[127] Z. Li, M. Shafiei, R. Ramamoorthi, K. Sunkavalli, and M. Chandraker. Inverse rendering for complex indoor scenes: shape, spatially-varying lighting and svbrdf from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2475–2484, 2020 (cited on p. 67).

[128] Z. Li, Z. Xu, R. Ramamoorthi, K. Sunkavalli, and M. Chandraker. Learning to reconstruct shape and spatially-varying reflectance from a single image. *ACM Transactions on Graphics (TOG)*, 37(6):1–11, 2018 (cited on p. 67).

[129] Z. Li, Q. Zheng, B. Shi, G. Pan, and X. Jiang. Dani-net: uncalibrated photometric stereo by differentiable shadow handling, anisotropic reflectance modeling, and neural inverse rendering. *arXiv preprint arXiv:2303.15101*, 2023 (cited on p. 128).

[130] T. Y. Lim, R. A. Yeh, Y. Xu, M. N. Do, and M. Hasegawa-Johnson. Time-frequency networks for audio super-resolution. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 646–650. IEEE, 2018 (cited on p. 14).

[131] P.-L. Lions, E. Rouy, and A. Tourin. Shape-from-shading, viscosity solutions and edges. *Numerische Mathematik*, 64:323–353, 1993 (cited on p. 43).

[132] F. Logothetis, R. Mecca, and R. Cipolla. A differential volumetric approach to multi-view photometric stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1052–1061, 2019 (cited on p. 45).

[133] F. Logothetis, R. Mecca, and R. Cipolla. Semi-calibrated near field photometric stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 941–950, 2017 (cited on pp. 39, 48).

[134] S. Lombardi, J. Saragih, T. Simon, and Y. Sheikh. Deep appearance models for face rendering. *ACM Transactions on Graphics (ToG)*, 37(4):1–13, 2018 (cited on p. 67).

[135] F. Lu, X. Chen, I. Sato, and Y. Sato. Symps: brdf symmetry guided photometric stereo for shape and light source estimation. *IEEE transactions on pattern analysis and machine intelligence*, 40(1):221–234, 2017 (cited on p. 45).

[136] F. Lu, Y. Matsushita, I. Sato, T. Okabe, and Y. Sato. Uncalibrated photometric stereo for unknown isotropic reflectances. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1490–1497, 2013 (cited on p. 65).

[137] Z. Lu, Y.-W. Tai, F. Deng, M. Ben-Ezra, and M. S. Brown. A 3d imaging framework based on high-resolution photometric-stereo and low-resolution depth. *International journal of computer vision*, 102(1-3):18–32, 2013 (cited on pp. 16, 57).

[138] F. Luan, S. Zhao, K. Bala, and Z. Dong. Unified shape and svbrdf recovery using differentiable monte carlo rendering. In *Computer Graphics Forum*, volume 40 of number 4, pages 101–113. Wiley Online Library, 2021 (cited on pp. 20, 40, 45).

[139] Y. Ma, S. Soatto, J. Košecká, and S. Sastry. *An invitation to 3-d vision: from images to geometric models*, volume 26. Springer, 2004 (cited on pp. 8, 10).

[140] R. Maier, K. Kim, D. Cremers, J. Kautz, and M. Nießner. Intrinsic3d: high-quality 3d reconstruction by joint appearance and geometry optimization with spatially-varying lighting. In *Proceedings of the IEEE international conference on computer vision*, pages 3114–3122, 2017 (cited on pp. 33, 40, 130).

[141] A. Marquina and S. J. Osher. Image super-resolution by tv-regularization and bregman iteration. *Journal of Scientific Computing*, 37:367–382, 2008 (cited on p. 15).

[142] M. Maximov, L. Leal-Taixé, M. Fritz, and T. Ritschel. Deep appearance maps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8729–8738, 2019 (cited on p. 67).

[143] R. Mecca and M. Falcone. Uniqueness and approximation of a photometric shape-from-shading model. *SIAM Journal on Imaging Sciences*, 6(1):616–659, 2013 (cited on pp. 39, 48).

[144] R. Mecca and Y. Quéau. Unifying diffuse and specular reflections for the photometric stereo problem. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9. IEEE, 2016 (cited on pp. 39, 48).

[145] R. Mecca, E. Rodola, and D. Cremers. Realistic photometric stereo using partial differential irradiance equation ratios. *Computers & Graphics*, 51:8–16, 2015 (cited on pp. 39, 48).

[146] R. Mecca, A. Tankus, A. Wetzler, and A. M. Bruckstein. A direct differential approach to photometric stereo with perspective viewing. *SIAM Journal on Imaging Sciences*, 7(2):579–612, 2014 (cited on pp. 39, 48).

[147] R. Mecca, A. Wetzler, A. M. Bruckstein, and R. Kimmel. Near field photometric stereo with point light sources. *SIAM Journal on Imaging Sciences*, 7(4):2732–2770, 2014 (cited on pp. 39, 48).

[148] R. Mecca, A. Wetzler, R. Kimmel, and A. M. Bruckstein. Direct shape recovery from photometric stereo with shadows. In *2013 International Conference on 3D Vision-3DV 2013*, pages 382–389. IEEE, 2013 (cited on pp. 39, 48).

[149] J. Mélou, Y. Quéau, F. Castan, and J.-D. Durou. A splitting-based algorithm for multi-view stereopsis of textureless objects. In *Scale Space and Variational Methods in Computer Vision: 7th International Conference, SSVM 2019, Hofgeismar, Germany, June 30–July 4, 2019, Proceedings 7*, pages 51–63. Springer, 2019 (cited on pp. 16, 33, 39).

[150] J. Mélou, Y. Quéau, J.-D. Durou, F. Castan, and D. Cremers. Beyond multi-view stereo: shading-reflectance decomposition. In *Scale Space and Variational Methods in Computer Vision: 6th International Conference, SSVM 2017, Kolding, Denmark, June 4-8, 2017, Proceedings 6*, pages 694–705. Springer, 2017 (cited on p. 33).

[151] J. Mélou, Y. Quéau, J.-D. Durou, F. Castan, and D. Cremers. Variational reflectance estimation from multi-view images. *Journal of Mathematical Imaging and Vision*, 60:1527–1546, 2018 (cited on p. 33).

[152] B. Mildenhall, P. Hedman, R. Martin-Brualla, P. P. Srinivasan, and J. T. Barron. Nerf in the dark: high dynamic range view synthesis from noisy raw images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16190–16199, 2022 (cited on pp. 19, 128).

[153] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos. Image segmentation using deep learning: a survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3523–3542, 2021 (cited on p. 67).

[154] Z. Mo, B. Shi, F. Lu, S.-K. Yeung, and Y. Matsushita. Uncalibrated Photometric Stereo Under Natural Illumination. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2936–2945, 2018 (cited on pp. 50, 63, 65, 74).

[155] T. Möllenhoff and D. Cremers. Sublabel-accurate discretization of nonconvex free-discontinuity problems. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1183–1191, 2017 (cited on p. 128).

[156] D. B. Mumford and J. Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *Communications on pure and applied mathematics*, 1989 (cited on p. 51).

[157] G. Nam, J. H. Lee, D. Gutierrez, and M. H. Kim. Practical svbrdf acquisition of 3d objects with unstructured flash photography. *ACM Transactions on Graphics (TOG)*, 37(6):1–12, 2018 (cited on p. 45).

[158] 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. Ieee, 2011 (cited on pp. 4, 11).

[159] H. Nguyen. *Gpu Gems 3*. Addison-Wesley Professional, first edition, 2007. ISBN: 9780321545428 (cited on p. 20).

[160] F. E. Nicodemus. Directional reflectance and emissivity of an opaque surface. *Applied optics*, 4(7):767–775, 1965 (cited on p. 21).

[161] C. Nieuwenhuis and D. Cremers. Spatially Varying Color Distributions for Interactive Multilabel Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35(5):1234–1247, 2013 (cited on pp. 66, 67, 74).

[162] M. Nimier-David, Z. Dong, W. Jakob, and A. Kaplanyan. Material and lighting reconstruction for complex indoor scenes with texture-space differentiable rendering, 2021 (cited on pp. 68–70, 128).

[163] M. Nimier-David, D. Vicini, T. Zeltner, and W. Jakob. Mitsuba 2: a retargetable forward and inverse renderer. *ACM Transactions on Graphics (TOG)*, 38(6):1–17, 2019 (cited on pp. 69, 128).

[164] R. Or - El, G. Rosman, A. Wetzler, R. Kimmel, and A. M. Bruckstein. RGBD-fusion: Real-time high precision depth recovery. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5407–5416, 2015 (cited on pp. 4, 33, 39, 43, 57, 60–62, 64, 73).

[165] M. Oren and S. K. Nayar. Generalization of the lambertian model and implications for machine vision. *International Journal of Computer Vision*, 14:227–251, 1995 (cited on p. 23).

[166] S. Osher and J. A. Sethian. Fronts propagating with curvature-dependent speed: algorithms based on hamilton-jacobi formulations. *Journal of computational physics*, 79(1):12–49, 1988 (cited on p. 51).

[167] M. R. Oswald, E. Töppe, and D. Cremers. Fast and globally optimal single view reconstruction of curved objects. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 534–541. IEEE, 2012 (cited on pp. 39, 66, 126).

[168] G. Oxholm and K. Nishino. Multiview shape and reflectance from natural illumination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2155–2162, 2014 (cited on p. 45).

[169] G. Oxholm and K. Nishino. Shape and reflectance from natural illumination. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part I 12*, pages 528–541. Springer, 2012 (cited on p. 43).

[170] T. Papadhimitri and P. Favaro. A new perspective on uncalibrated photometric stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1474–1481, 2013 (cited on pp. 40, 46).

[171] T. Papadhimitri and P. Favaro. Uncalibrated near-light photometric stereo, 2014 (cited on pp. 40, 47).

[172] J. Park, H. Kim, Y.-W. Tai, M. S. Brown, and I. S. Kweon. High-quality depth map upsampling and completion for rgb-d cameras. *IEEE Transactions on Image Processing*, 23(12):5559–5572, 2014 (cited on pp. 16, 55–57).

[173] J. Park, H. Kim, Y.-W. Tai, M. S. Brown, and I. Kweon. High quality depth map upsampling for 3d-tof cameras. In *2011 International Conference on Computer Vision*, pages 1623–1630. IEEE, 2011 (cited on pp. 16, 55–57).

[174] J. Park, S. N. Sinha, Y. Matsushita, Y.-W. Tai, and I. S. Kweon. Robust multiview photometric stereo using planar mesh parameterization. *IEEE transactions on pattern analysis and machine intelligence*, 39(8):1591–1604, 2016 (cited on p. 45).

[175] S. G. Parker, J. Bigler, A. Dietrich, H. Friedrich, J. Hoberock, D. Luebke, D. McAllister, M. McGuire, K. Morley, A. Robison, and M. Stich. Optix: a general purpose ray tracing engine. *ACM Trans. Graph.*, 29(4), July 2010. ISSN: 0730-0301. DOI: 10.1145/1778765.1778803 (cited on p. 20).

[176] M. Pharr and R. Fernando. *GPU Gems 2: Programming Techniques for High-Performance Graphics and General-Purpose Computation (Gpu Gems)*. Addison-Wesley Professional, 2005. ISBN: 0321335597 (cited on p. 20).

[177] M. Pharr, W. Jakob, and G. Humphreys. *Physically based rendering: From theory to implementation*. Morgan Kaufmann, 2016 (cited on p. 20).

[178] B. T. Phong. Illumination for computer generated pictures. *Communications of the ACM*, 18(6):311–317, 1975 (cited on pp. 21, 22).

[179] H. Poincaré. *La dynamique de l'électron*. A. Dumas, 1913 (cited on p. 47).

[180] E. Prados and O. D. Faugeras. "Perspective shape from shading"and viscosity solutions. In *ICCV*, volume 3, page 826, 2003 (cited on p. 43).

[181] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press, 2007 (cited on p. 32).

[182] Y. Quéau, R. Bruneau, J. Mélou, J.-D. Durou, and F. Lauze. On photometric stereo in the presence of a refractive interface. In *9th International Conference on Scale Space and Variational Methods in Computer Vision (SSVM 2023)*, pages 1–13, 2023 (cited on p. 39).

[183] Y. Quéau, B. Durix, T. Wu, D. Cremers, F. Lauze, and J.-D. Durou. Led-based photometric stereo: modeling, calibration and numerical solution. *Journal of Mathematical Imaging and Vision*, 60:313–340, 2018 (cited on p. 39).

[184] Y. Quéau, J.-D. Durou, and J.-F. Aujol. Normal integration: a survey. *Journal of Mathematical Imaging and Vision*, 60:576–593, 2018 (cited on p. 39).

[185] Y. Quéau, J.-D. Durou, and J.-F. Aujol. Variational methods for normal integration. *Journal of Mathematical Imaging and Vision*, 60:609–632, 2018 (cited on p. 39).

[186] Y. Quéau, F. Lauze, and J.-D. Durou. A-tv algorithm for robust perspective photometric stereo with spatially-varying lightings. In *Scale Space and Variational Methods in Computer Vision: 5th International Conference, SSVM 2015, Lège-Cap Ferret, France, May 31-June 4, 2015, Proceedings*, pages 498–510. Springer, 2015 (cited on p. 48).

[187] Y. Quéau, R. Mecca, and J.-D. Durou. Unbiased photometric stereo for colored surfaces: a variational approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4359–4368, 2016 (cited on pp. 39, 48, 49, 64, 67).

[188] Y. Quéau, J. Mélou, F. Castan, D. Cremers, and J.-D. Durou. A variational approach to shape-from-shading under natural illumination. In *Energy Minimization Methods in Computer Vision and Pattern Recognition: 11th International Conference, EMMCVPR 2017, Venice, Italy, October 30–November 1, 2017, Revised Selected Papers 11*, pages 342–357. Springer, 2018 (cited on pp. 16, 33, 39, 43, 57, 61, 62).

[189] Y. Quéau, J. Mélou, J.-D. Durou, and D. Cremers. Dense multi-view 3d-reconstruction without dense correspondences. *arXiv preprint arXiv:1704.00337*, 2017 (cited on p. 39).

[190] Y. Quéau, T. Wu, F. Lauze, J.-D. Durou, and D. Cremers. A non-convex variational approach to photometric stereo under inaccurate lighting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 99–108, 2017 (cited on pp. 23, 39, 45, 66).

[191] R. Ramamoorthi and P. Hanrahan. An efficient representation for irradiance environment maps. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 497–500, 2001 (cited on pp. 28, 30, 49).

[192] R. Ranftl, S. Gehrig, T. Pock, and H. Bischof. Pushing the limits of stereo using variational stereo estimation. In *2012 IEEE Intelligent Vehicles Symposium*, pages 401–407. IEEE, 2012 (cited on p. 57).

[193] S. R. Richter and S. Roth. Discriminative shape from shading in uncalibrated illumination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1128–1136, 2015 (cited on p. 43).

[194] J. Richter-Gebert. *Perspectives on projective geometry: a guided tour through real and complex geometry*. Springer, 2011 (cited on p. 10).

[195] E. Rouy and A. Tourin. A viscosity solutions approach to shape-from-shading. *SIAM Journal on Numerical Analysis*, 29(3):867–884, 1992 (cited on p. 43).

[196] J. Salvi, X. Armangué, and J. Batlle. A comparative review of camera calibrating methods with accuracy evaluation. *Pattern recognition*, 35(7):1617–1635, 2002 (cited on p. 11).

[197] J. Salvi, S. Fernandez, T. Pribanic, and X. Llado. A state of the art in structured light patterns for surface profilometry. *Pattern recognition*, 43(8):2666–2680, 2010 (cited on p. 11).

[198] S. Sang and M. Chandraker. Single-shot neural relighting and svbrdf estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16*, pages 85–101. Springer, 2020 (cited on p. 67).

[199] H. Santo, M. Samejima, Y. Sugano, B. Shi, and Y. Matsushita. Deep photometric stereo network. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 501–509, 2017 (cited on p. 128).

[200] H. Santo, M. Samejima, Y. Sugano, B. Shi, and Y. Matsushita. Deep photometric stereo networks for determining surface normal and reflectances. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):114–128, 2020 (cited on p. 128).

[201] D. Scharstein and R. Szeliski. High-accuracy stereo depth maps using structured light. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.* Volume 1, pages I–I. IEEE, 2003 (cited on p. 11).

[202] H. Schieber, F. Deuser, B. Egger, N. Oswald, and D. Roth. Nerftrinsic four: an end-to-end trainable nerf jointly optimizing diverse intrinsic and extrinsic camera parameters. *arXiv preprint arXiv:2303.09412*, 2023 (cited on p. 128).

[203] C. Schlick. An inexpensive brdf model for physically-based rendering. In *Computer graphics forum*, volume 13 of number 3, pages 233–246. Wiley Online Library, 1994 (cited on pp. 25, 26).

[204] C. Schmitt, S. Donne, G. Riegler, V. Koltun, and A. Geiger. On joint estimation of pose, geometry and svbrdf from a handheld scanner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3493–3503, 2020 (cited on p. 67).

[205] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 1, pages 519–528. IEEE, 2006 (cited on p. 4).

[206] S. A. Shafer. Using color to separate reflection components. *Color Research & Application*, 10(4):210–218, 1985 (cited on p. 21).

[207] B. Shi, K. Inose, Y. Matsushita, P. Tan, S.-K. Yeung, and K. Ikeuchi. Photometric stereo using internet images. In *2014 2nd International Conference on 3D Vision*, volume 1, pages 361–368. IEEE, 2014 (cited on pp. 50, 63–65).

[208] B. Shi, Z. Wu, Z. Mo, D. Duan, S.-K. Yeung, and P. Tan. A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3707–3716, 2016 (cited on pp. 4, 45, 53, 66).

[209] Y. Shirai and M. Suwa. Recognition of polyhedrons with a range finder. In *IJCAI*, pages 80–87, 1971 (cited on p. 11).

[210] B. Smith. Geometrical shadowing of a random rough surface. *IEEE transactions on antennas and propagation*, 15(5):668–671, 1967 (cited on p. 25).

[211] W. Smith and F. Fang. Height from photometric ratio with model-based light source selection. *Computer Vision and Image Understanding*, 145:128–138, 2016 (cited on pp. 39, 48).

[212] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: exploring photo collections in 3d. In *ACM siggraph 2006 papers*, pages 835–846. 2006 (cited on p. 64).

[213] K. Sofiiuk, I. A. Petrov, and A. Konushin. Reviving iterative training with mask guidance for interactive segmentation. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3141–3145. IEEE, 2022 (cited on pp. 66, 67).

[214] G. Song, H. Myeong, and K. M. Lee. Seednet: automatic seed generation with deep reinforcement learning for robust interactive segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1760–1768, 2018 (cited on pp. 66, 67).

[215] J. Steurer. Tri-focal rig (practical camera configurations for image and depth acquisition). In *SMPTE 2013 Annual Technical Conference & Exhibition*, pages 1–15. SMPTE, 2013 (cited on p. 127).

[216] E. Strekalovskiy and D. Cremers. Real-time minimization of the piecewise smooth mumford-shah functional. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II 13*, pages 127–141. Springer, 2014 (cited on p. 130).

[217] J. Stühmer, S. Gumhold, and D. Cremers. Real-time dense geometry from a handheld camera. In *Pattern Recognition: 32nd DAGM Symposium, Darmstadt, Germany, September 22-24, 2010. Proceedings 32*, pages 11–20. Springer, 2010 (cited on p. 11).

[218] J. Sun, M. Smith, L. Smith, S. Midha, and J. Bamber. Object surface recovery using a multi-light photometric stereo technique for non-lambertian surfaces subject to shadows and specularities. *Image and Vision Computing*, 25(7):1050–1057, 2007 (cited on p. 45).

[219] A. Tankus, N. Sochen, and Y. Yeshurun. Shape-from-shading under perspective projection. *International Journal of Computer Vision*, 63:21–43, 2005 (cited on p. 43).

[220] The GIMP Development Team. GIMP. www.gimp.org, version 2.8.22, June 12, 2019 (cited on pp. 66, 74).

[221] K. E. Torrance and E. M. Sparrow. Theory for off-specular reflection from roughened surfaces. *Josa*, 57(9):1105–1114, 1967 (cited on pp. 22, 24, 129).

[222] T. Trowbridge and K. P. Reitz. Average irregularity representation of a rough surface for ray reflection. *JOSA*, 65(5):531–536, 1975 (cited on pp. 25, 26, 129).

[223] M. Unger, T. Pock, M. Werlberger, and H. Bischof. A convex approach for variational super-resolution. In *Pattern Recognition: 32nd DAGM Symposium, Darmstadt, Germany, September 22-24, 2010. Proceedings 32*, pages 313–322. Springer, 2010 (cited on pp. 14, 16).

[224] V. Usenko, N. Demmel, and D. Cremers. The double sphere camera model. In *2018 International Conference on 3D Vision (3DV)*, pages 552–560. IEEE, 2018 (cited on p. 8).

[225] J. Van Ouwerkerk. Image super-resolution survey. *Image and vision Computing*, 24(10):1039–1052, 2006 (cited on p. 14).

[226] E. Veach. *Robust Monte Carlo methods for light transport simulation*. Stanford University, 1998 (cited on p. 68).

[227] D. Vlasic, P. Peers, I. Baran, P. Debevec, J. Popović, S. Rusinkiewicz, and W. Matusik. Dynamic shape capture using multi-view photometric stereo. In *ACM SIGGRAPH Asia 2009 papers*, pages 1–11. 2009 (cited on p. 45).

[228] O. Vogel, A. Bruhn, J. Weickert, and S. Didas. Direct shape-from-shading with adaptive higher order regularisation. In *Scale Space and Variational Methods in Computer Vision: First International Conference, SSVM 2007, Ischia, Italy, May 30-June 2, 2007. Proceedings 1*, pages 871–882. Springer, 2007 (cited on p. 43).

[229] L. von Stumberg, V. Usenko, J. Engel, J. Stueckler, and D. Cremers. From monocular SLAM to autonomous drone exploration. In *European Conference on Mobile Robots (ECMR)*, Sept. 2017 (cited on p. 3).

[230] B. Walter, S. R. Marschner, H. Li, and K. E. Torrance. Microfacet models for refraction through rough surfaces. In *Proceedings of the 18th Eurographics conference on Rendering Techniques*, pages 195–206, 2007 (cited on pp. 25, 26, 129).

[231] P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang. Neus: learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021 (cited on pp. 128, 129).

[232] Z. Wang, S. Wu, W. Xie, M. Chen, and V. A. Prisacariu. Nerf−−: neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021 (cited on pp. 40, 128).

[233] S. Wanner and B. Goldluecke. Spatial and angular variational super-resolution of 4d light fields. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pages 608–621. Springer, 2012 (cited on p. 14).

[234] G. J. Ward. Measuring and modeling anisotropic reflection. In *Proceedings of the 19th annual conference on Computer graphics and interactive techniques*, pages 265–272, 1992 (cited on p. 22).

[235] O. Wasenmüller and D. Stricker. Comparison of kinect v1 and v2 depth images in terms of accuracy and precision. In *Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13*, pages 34–45. Springer, 2017 (cited on p. 13).

[236] M. Werlberger, W. Trobin, T. Pock, A. Wedel, D. Cremers, and H. Bischof. Anisotropic huber-l1 optical flow. In *BMVC*, volume 1 of number 2, page 3, 2009 (cited on pp. 16, 57).

[237] P. M. Will and K. S. Pennington. Grid coding: a preprocessing technique for robot and machine vision. *Artificial Intelligence*, 2(3-4):319–329, 1971 (cited on p. 11).

[238] R. J. Woodham. Photometric method for determining surface orientation from multiple images. *Optical engineering*, 19(1):139–144, 1980 (cited on pp. 4, 44, 66, 128).

[239] C. Wu, M. Zollhöfer, M. Nießner, M. Stamminger, S. Izadi, and C. Theobalt. Real-time shading-based refinement for consumer depth cameras. *ACM Transactions on Graphics (ToG)*, 33(6):1–10, 2014 (cited on pp. 39, 43, 57, 59–62).

[240] L. Wu, A. Ganesh, B. Shi, Y. Matsushita, Y. Wang, and Y. Ma. Robust photometric stereo via low-rank matrix completion and recovery. In *Computer Vision–ACCV 2010: 10th Asian Conference on Computer Vision, Queenstown, New Zealand, November 8-12, 2010, Revised Selected Papers, Part III 10*, pages 703–717. Springer, 2011 (cited on pp. 45, 47).

[241] X. Xiang, Y. Tian, Y. Zhang, Y. Fu, J. P. Allebach, and C. Xu. Zooming slow-mo: fast and accurate one-stage space-time video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3370–3379, 2020 (cited on p. 14).

[242] J. Xie, R. S. Feris, and M.-T. Sun. Edge guided single depth image super resolution. In *IEEE International Conference on Image Processing (ICIP)*, pages 3773–37777, 2014 (cited on p. 73).

[243] Z. Xu, K. Sunkavalli, S. Hadap, and R. Ramamoorthi. Deep image-based relighting from optimal sparse samples. *ACM Transactions on Graphics (ToG)*, 37(4):1–13, 2018 (cited on p. 67).

[244] Z. Xu, R. Schwarte, H.-G. Heinol, B. Buxbaum, and T. Ringbeck. Smart pixel: photonic mixer device (pmd); new system concept of a 3d-imaging camera-on-a-chip, 1998 (cited on p. 12).

[245] J. Yang, J. Wright, T. S. Huang, and Y. Ma. Image super-resolution via sparse representation. *IEEE transactions on image processing*, 19(11):2861–2873, 2010 (cited on p. 14).

[246] Q. Yang, R. Yang, J. Davis, and D. Nister. Spatial-Depth Super Resolution for Range Images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2007 (cited on pp. 16, 55–57, 73).

[247] W. Yang, G. Chen, C. Chen, Z. Chen, and K.-Y. K. Wong. Ps-nerf: neural inverse rendering for multi-view photometric stereo. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part I*, pages 266–284. Springer, 2022 (cited on p. 45).

[248] Y. Yao, J. Zhang, J. Liu, Y. Qu, T. Fang, D. McKinnon, Y. Tsin, and L. Quan. Neilf: neural incident light field for physically-based material estimation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI*, pages 700–716. Springer, 2022 (cited on pp. 128, 129).

[249] L. Yariv, J. Gu, Y. Kasten, and Y. Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34:4805–4815, 2021 (cited on pp. 128, 129).

[250] L.-F. Yu, S.-K. Yeung, Y.-W. Tai, and S. Lin. Shading-based shape refinement of rgb-d images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1415–1422, 2013 (cited on pp. 39, 43, 57–62).

[251] L. Yue, H. Shen, J. Li, Q. Yuan, H. Zhang, and L. Zhang. Image super-resolution: the techniques, applications, and future. *Signal processing*, 128:389–408, 2016 (cited on p. 14).

[252] A. Yuille and D. Snow. Shape and albedo from multiple images using integrability. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 158–164. IEEE, 1997 (cited on pp. 4, 40, 46).

[253] E. Zhang, M. F. Cohen, and B. Curless. Emptying, refurnishing, and relighting indoor spaces. *ACM Transactions on Graphics (TOG)*, 35(6):1–14, 2016 (cited on pp. 68, 70).

[254] J. Zhang, Y. Yao, S. Li, J. Liu, T. Fang, D. McKinnon, Y. Tsin, and L. Quan. Neilf++: inter-reflectable light fields for geometry and material estimation. *arXiv preprint arXiv:2303.17147*, 2023 (cited on p. 129).

[255] K. Zhang, F. Luan, Z. Li, and N. Snavely. Iron: inverse rendering by optimizing neural sdfs and materials from photometric images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5565–5574, 2022 (cited on pp. 20, 40, 45).

[256] K. Zhang, F. Luan, Q. Wang, K. Bala, and N. Snavely. Physg: inverse rendering with spherical gaussians for physics-based material editing and relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5453–5462, 2021 (cited on pp. 20, 40).

[257] R. Zhang, P.-S. Tsai, J. E. Cryer, and M. Shah. Shape-from-shading: a survey. *IEEE transactions on pattern analysis and machine intelligence*, 21(8):690–706, 1999 (cited on p. 43).

[258] H.-K. Zhao, T. Chan, B. Merriman, and S. Osher. A variational level set approach to multiphase motion. *Journal of computational physics*, 127(1):179–195, 1996 (cited on p. 51).

[259] Q. Zheng, A. Kumar, B. Shi, and G. Pan. Numerical reflectance compensation for non-lambertian photometric stereo. *IEEE Transactions on Image Processing*, 28(7):3177–3191, 2019 (cited on p. 45).

[260] Z. Zhong, X. Liu, J. Jiang, D. Zhao, and X. Ji. Guided depth map super-resolution: a survey. *ACM Computing Surveys*, 2023 (cited on p. 16).

[261] Q.-Y. Zhou and V. Koltun. Color map optimization for 3d reconstruction with consumer depth cameras. *ACM Transactions on Graphics (ToG)*, 33(4):1–10, 2014 (cited on p. 4).

[262] Z. Zhou, Z. Wu, and P. Tan. Multi-view photometric stereo with spatially varying isotropic materials. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1482–1489, 2013 (cited on p. 45).

[263] M. Zollhöfer, A. Dai, M. Innmann, C. Wu, M. Stamminger, C. Theobalt, and M. Nießner. Shading-based refinement on volumetric signed distance functions. *ACM Transactions on Graphics (TOG)*, 34(4):1–14, 2015 (cited on pp. 4, 33, 130).

# About the Author

Björn Häfner holds a B.Sc. in Mathematics from OTH Regensburg (2013) and an M.Sc. in Mathematics in Science and Engineering from the Technical University of Munich (2016). In November 2016, he commenced his doctoral studies in the Chair for Computer Vision and Artificial Intelligence at the Technical University of Munich under the supervision of Prof. Dr. Daniel Cremers. During the course of his doctoral studies, he has visited the GREYC Image Group in Caen, France as a visiting Ph.D. student in May 2019, completed two research internships at Reality Labs Research (Meta) in 2019 and 2021, and worked as an external research collaborator at Pro Unlimited @ Meta, closely collaborating with Reality Labs Research (Meta) in 2020.

His research interests include RGB-D data processing for 3D reconstruction, variational methods, shape-from-shading, photometric stereo, inverse rendering, and super-resolution. While pursuing his Ph.D. degree, he has conducted research in several areas related to high-quality, photorealistic 3D reconstruction, which has resulted in 11 peer-reviewed publications in top-tier conferences and journals. As a teaching assistant at TUM, he has supervised various lab courses and lecture exercises and mentored a total of 8 student projects (Bachelor's theses, Master's theses, etc.). Additionally, he has served as a reviewer for prestigious conferences and journals such as CVPR, ICCV, ECCV, and Springer.