Shape, Reflectance and Illumination Estimation From Mobile Devices

Submitted in partial fulfillment of the requirements for the

degree of

Doctor of Philosophy

in

Department of Electrical and Computer Engineering

Zhuo Hui

B.S., Electronics and Information Engineering
The Hong Kong Polytechnic University
M.S., Electrical and Computer Engineering
Carnegie Mellon University

Carnegie Mellon University
Pittsburgh, PA

January, 2019

Abstract

Inferring shape, reflectance and illumination from an image enables us to better understand a scene and has opened up a wide range of socially-compelling applications ranging from virtual reality, entertainment and home surveillance. However, this is a particularly challenging problem since all the factors are combined into a single observation. At the same time, the ubiquity of mobile devices has raised immense opportunities for wide spread adoption of such techniques. Toward this end, this dissertation addresses the problem for the shape, reflectance and illumination estimation using sensors and illuminants commonly found on commodity devices, such as smart phones and tablets. We focus on two subsets of the problem. We first address the problem of estimating the shape and reflectance of objects that exhibit spatially-varying reflectance under *known* single illuminant. Our contributions here are two-fold. First, we provide a non-iterative technique for per-pixel shape and reflectance, that is able to outperform the state-of-the-art methods on a wide range of real scenes. Second, we translate our technique to mobile de-

vices and demonstrate capabilities in estimating as well as editing reflectance in spite of the flash unit and camera sensor being collocated. Next, for a Lambertian scene being illuminated with multiple light sources, we propose a method to separate and manipulate the scene illuminants based on their spectral differences. As before, we make two contributions. First, we derive physics-based constraints for the flash/no-flash image pairs and provide identifiable analysis with respect to the number of light sources in the scene. We show that this separation can be used to support applications like white balancing, lighting editing, and RGB photometric stereo, where we demonstrate results that outperform state-of-the-art techniques on a wide range of images. Second, to address the limitations of the flash light for the mobile devices, such as the presence of strong ambient light as well as the scenes with large depth variations, we further demonstrate the ability to separate the image by simulating the flash image with a deep neutral network. We show that we are able to produce high-quality outputs that match the performance of previous methods that required a flash/no-flash pair, while being more practical in requiring only a single image. We believe that, all together, the techniques developed in this dissertation makes a significant advance in our ability to not just estimate a scene's shape, reflectance, and illumination but also enable subsequent inference and post-processing capabilities.

Acknowledgements

I would like to express my sincerest gratitude to my advisor Prof Aswin Sankaranarayanan for the continuous support of my Ph.D study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D study.

I am very grateful to Dr Kalyan Sunkavalli, who provided me an opportunity to join Adobe as intern as well as consistently provided the help, encouragement and guidance in my graduate studies. Many interesting ideas and research work have shaped from my internships at Adobe.

I would like to thank the rest of my thesis committee: Prof. Vijayakumar Bhagavatula and Prof. Ioannis Gkioulekas, for their insightful comments and encouragement, but also for the hard questions which instill me to widen my research from various perspectives.

I thank fellow lab-mates, Jian, Rick, Chia-Yin, Vishwa, Yi and Byeongjoo, for the stimulating discussions, for sleepless nights we were working together before deadlines, and for all the fun we have had in the past few years. Also, I am grateful to my friends, Ke and Zhiding for providing company for the lunch and dinner, to Cheng and Qi for the suggestions in the career plan.

I would like to express my sincerest appreciation to my family. My parents Shuli and Lijuan provide continuous support and encouragement to my study as well as being my life-coach to shape my virtues. A very special gratitude goes out to my wife, Meng, for always being there when I need the support and company to against the difficulties in the Ph.D life. I would also thank all family members in my hometown, who provide their company to my parents when I was not around.

This dissertation was supported in part by the NSF grant CCF-1117939, the NSF CAREER grant CCF-1652569, the NGIA grant HM0476-17-1-2000, the ARO Grant W911NF-15-1-0126, a gift from Adobe Research, and the CIT Dean's Fellowship.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Scene understanding, a central problem in computer vision, refers to the ability to describe a scene in terms of the shape of objects present in the scene, the nature of materials that comprise these objects, and the size, color and orientation of lighting incident upon them. Over the last few decades, advances in inferring such scene properties have opened up a wide range of socially-compelling applications in virtual reality, entertainment, home monitoring, and robotics. While this problem seems trivial to us since we perform it easily and often subconsciously, for computers, it is an extremely challenging task since they perceive the images in a completely different way. In this dissertation, we focus on the ability to infer the scene properties, including shape, reflectance and illumination, and in particular, use it for post-process manipulation of reflectance and illumination in a scene. Further, we also focus on enabling such capabilities under relatively unconstrained conditions and using commodity hardware that are commonly found on smart phones and tablets.

To achieve these capabilities in scene understanding, we first have to understand how reflectance, shape and illumination are encoded in image measurements of a scene. The intensity observed at a pixel in a image is a complex function of scene geometry, materials properties, illumination, the imaging device, and subsequent post-processing. To disentangle each of these factors is a highly ill-posed problem. Given the complexity of the general inverse problem, two important subsets have been studied in detail in prior literature: first, *estimating the shape and reflectance* by assuming that illumination of the scene — in terms of color and orientation — is fully characterized, and second, *estimating the illumination* by imposing strong prior on the reflectance (e.g. Lambertian material).

To estimate the shape and reflectance under known illumination, scenes with Lambertian reflectance have been studied extensively in the context of photometric stereo [76, 149, 152] due to the immense simplification that such an assumption provides. Unfortunately, real-life scenes often involve non-Lambertian materials that interact with light in complex ways; this creates a significant disconnect between theory and practice. While there have been some efforts [9, 63, 93] for developing photometric stereo for spatially-varying reflectance, they either rely on restrictive assumptions on the materials or extra reference materials in the scene [71], both of which make them impractical in real world scenarios. In addition, most of these approaches rely on dense sampling of the reflectance, thereby requiring precisely calibrated, often prohibitive expensive acquisition setup, which hinder the accessibility for the average consumer. While photometric stereo has been studied in the uncalibrated setting, where the orientation of the light sources are unknown, the performance in the absence of calibration is often significantly worse than their calibrated counterpart [138].

In parallel, many techniques have been developed to explicitly estimate the illumination of the scene, rather than assuming it can be known. Color constancy [54, 55], or white balancing, — the problem of correcting for the illuminant spectrum — is a closely related light estimation problem, and has been extensively studied in the literature. Specifically, the goal is to render a new photograph such that the observed color of scene points is influenced only by their reflectance and not by the spectral profile on the scene illumination. As is to be expected, color constancy is an integral part of most digital camera pipelines as well as image processing tools. While there are numerous techniques for performing color constancy, most of them are not able to handle multiple illuminants.

This dissertation addresses the problems for shape, reflectance and illumination estimation using mobile devices for scenes that have complex spatially-varying reflectance and illumination. Solving these problems in a mobile setting raises some additional challenges, in addition to those already enumerated above. As compared to carefully calibrated light stages, there is a significant reduction in our ability to both control the illumination in the scene as well as additional challenges in terms of lack of extrinsic camera calibration. Specifically, many of the factors, such as lighting direction, camera pose, that can be obtained via precisely calibration with light stages, become unknown and need to be estimated, which increases the complexity of the problem. This poses a significant challenge that we seek to address in this dissertation.

The goal of our work is to address two subset of the problems: first, for the scene under single illuminant, we enable the capability in manipulating the reflectance of the objects by solving for the shape and reflectance of the objects that exhibit spatially varying BRDF (SV-BRDF) (see Figure 1.1), and second, for the scene under mixture of illuminants, we demonstrate the capability in manipulating the scene illumination by separating the captured photograph into several images, each of which only lit by a single light source(see Figure 1.2). That is, we aim for solving the problems defined as:

Shape and reflectance estimation for the object exhibiting non-Lambertian spatially varying BRDF by the use of mobile devices. We show that the performance of our technique on a wide range of simulated and real scenes where we outperform competing methods. An example is shown in Figure 1.1.

Illumination decomposition for the image captured under the mixture of light sources (see Figure 1.2). We demonstrate that this separation can be used to support applications like white balancing, lighting editing, and RGB photometric stereo, where we demonstrate results that outperform state-of-the-art techniques on a wide range of images.

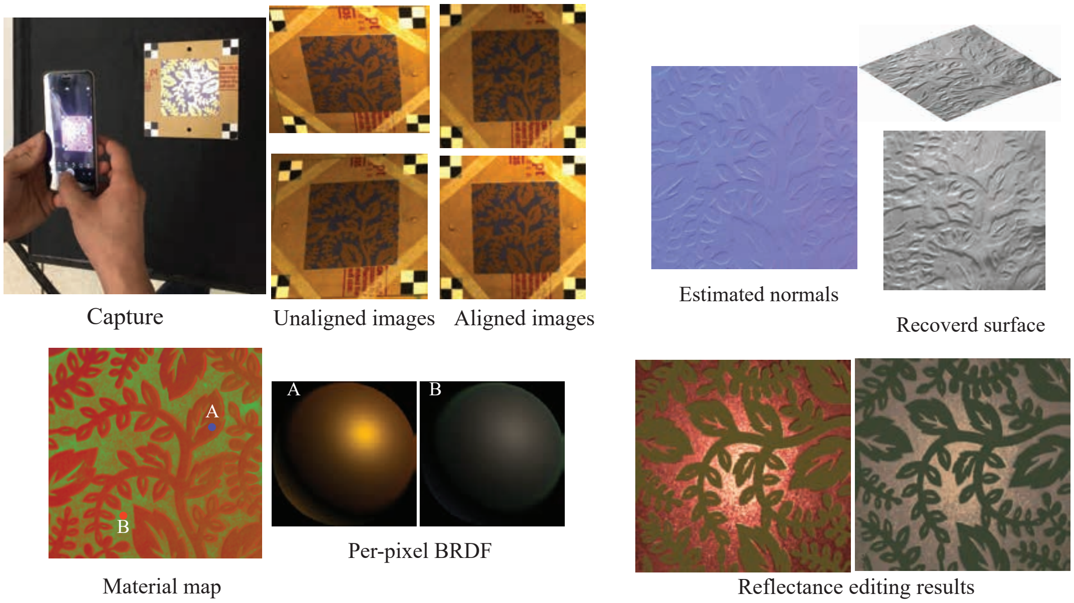
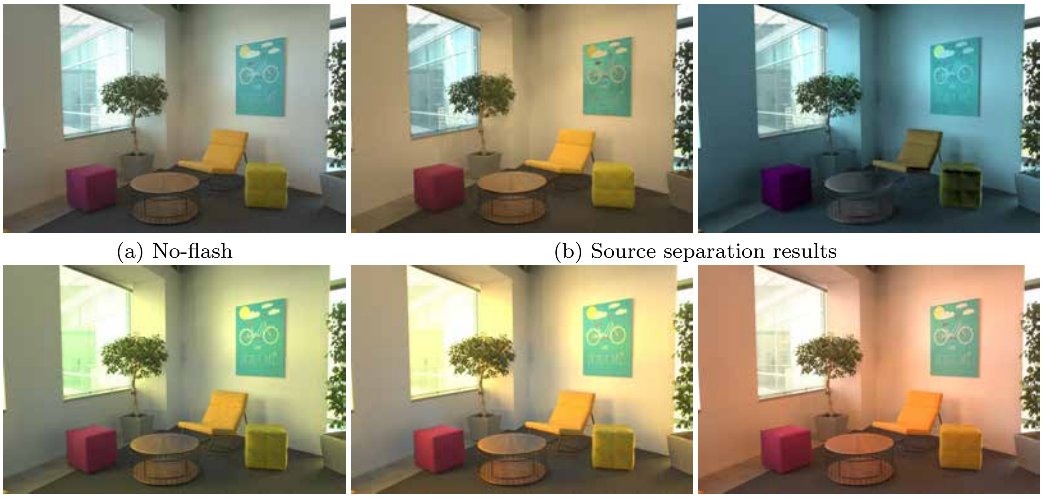


Figure 1.1: We acquire multiple images of a near-planar object using the camera and the flash unit on a mobile phone and subsequently, estimate the surface normals as well as the spatially-varying BRDF at each pixel.

For both problems, we need to solve for particularly challenging problems to infer the scene properties. Before we discuss the specific technical contributions in detail, we first go over the challenges underlying the problems solved in this dissertation as well as discuss key related work.

1.1 Prior work

Reconstructing a scene — and all its properties including shape, reflectance of the material, and ambient illumination — from portable devices is one of the long-standing goals of computer vision and graphics, and has been extensively



(a) No-flash

(b) Source separation results

(c) Illumination editing results

Figure 1.2: The scene in (a) is lit by cool sky illumination from the window on the left and warm indoor lighting from the top. Given a pair of no-flash/flash images, our method separates the no-flash image into two images lit by each of these illuminants (b) and estimates their spectral distribution (insets in (b)). Using our illuminant estimates, we are able to edit the illumination in the photograph (c) by changing the individual spectra of the light sources (insets in (c)).

studied in the literature. However, this is a highly ill-posed problem since all these properties are intrinsically tied to each other, estimating each of them often relies on reconstructing the others. In this section, we will specifically focus on the challenges for recovering SV-BRDF and shape from controlled illuminants, as well as predicting the ambient illumination colors.

1.1.1 SV-BRDF and shape estimation.

Reflectance properties play an important role in the appearance of objects in a scene. For an opaque object, these properties are represented by the 4-

Acquisition	Method	Material	SV-BRDF	Camera	Illumination
Uncalibrated	Lambertian	Yes	SfM [154]	Moving	Ambient
	SL [105]	Lambertian	Yes	Moving	Structured
	PS [11]	Lambertian	Yes	Static	Directional
	Intensity [100]	General	No	Static	Directional
Calibrated	PS [152]	Lambertian	Yes	Static	Directional
	Parametric [63]	General	Yes	Static	Directional
	Data-driven [103]	General	No	Static	Directional
	Proposed [78]	General	Yes	Moving	Point

Table 1.1: Shape and BRDF estimation methods for different camera and illumination setups. Note that SV-BRDF indicates whether the target object is spatially-varying or homogeneous, SfM refers structure from motion methods, SL denotes the methods relying on structured light, PS refers the photometric stereo. For the non-Lambertian methods, we refer parametric as the methods by using parametric model to characterize material BRDFs while data-driven refers to the methods relying on measured BRDFs database. For our method, we focus on spatially varying BRDFs by the use of moving camera with point light source, i.e. the setup for the mobile device.

D bidirectional reflectance distribution function (BRDF), which completely characterizes how a material interacts with incident light. However, measuring the BRDF of a material is since it often requires dense sampling of the 4-D space, hence requiring either large amount of input images or prior knowledge of the materials. To address this, techniques have been developed by incorporating additional prior on the BRDF as well as utilizing different camera and illuminant setup. We show a variety of shape and BRDF methods with respect to the assumption on the materials as well as the setup for the camera and illumination in Table 1.1.

Lambertian BRDF. The majority of past techniques rely on the Lambertian assumption, in which the BRDF of the object can be characterized as a constant. An example of this is classical photometric stereo [152], as shown in Figure 1.3, where the shape of the objects is able to be recovered by using three images under a fixed view-point and varying illumination.

In parallel, many techniques have been developed to recover the shape from photographs captured under the same illumination but varying camera poses [56, 57]. While significant progress has been made on the Lambertian materials, the real world consists of a variety of materials exhibiting non-Lambertian reflectance, i.e. shinny or mirror-like objects. This always leads to immense increase in the dimensionality of BRDF. In addition, the key challenge is that the reflectance, characterized in terms of BRDF, and the shape, characterized in terms of surface normals, are inherently coupled and need to be estimated jointly. Further, the SV-BRDF is a 6D function of space and incident/outgoing angles and hence, can be very high-dimensional. In the absence of additional assumptions, estimating the SV-BRDF requires a large number of input images for robust estimation.

Homogeneous BRDF. Instead of assuming Lambertian material, a common assumption for enabling computationally tractable model is that the object is made of the same material or all the pixels share the same BRDF. This provides a significant reduction in the dimensionality of BRDF since we are able to pool the information together across pixels. However, few real

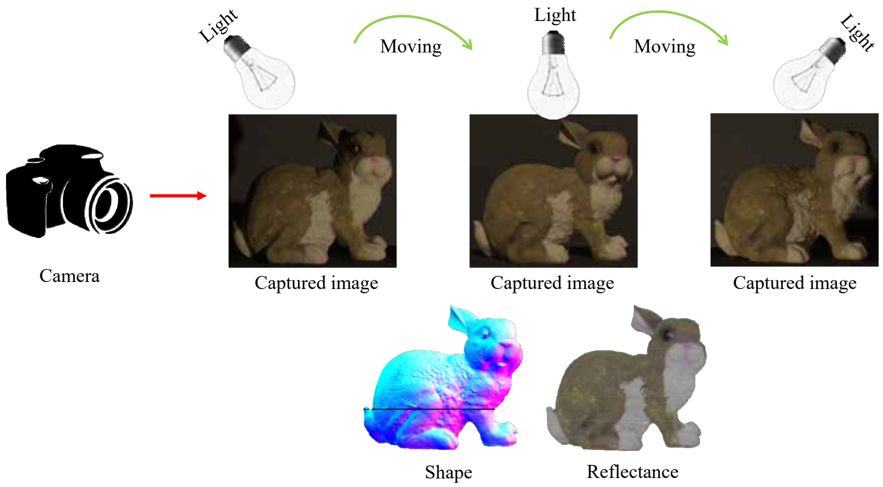


Figure 1.3: Given the static camera and moving calibrated light source, it is suffice to recover the shape and reflectance for the object with Lambertian material with three images.

objects are exactly made of single material and BRDF may vary even for adjacent pixels. To this end, bulk of attentions are received for the problem defined in terms of the spatially varying BRDF.

Spatially-varying BRDF. To estimate the spatially-varying BRDF, it always requires large amount of input images. To make the problem computational tractable, the assumptions are made on the BRDF to restrict the underlying solution space. The initial attempt is made by modeling the BRDF with a variety of physical-based parametric models [22,39,150]. While it provides immense reduction in the dimensionality, i.e. from millions of unknowns to less than 10, it is inherently limited by the generalizabity to the complex materials with highly specular lobes [107]. To address this, Matusik et al. [103] actually measure a wide range of material BRDFs and introduce the data-driven model to characterize the target unkown reflectance. In particular, Lawrence et al. [93] assume that BRDF at each pixel is characterized as a weighted combination of a few, unknown reference BRDFs. The BRDF is now represented using the reference BRDFs and their relative abundances at each pixel. The problem of shape and BRDF estimation reduces to alternating minimization over the surface normals, the reference BRDFs, and abundances of the reference BRDFs at each pixel. However, to solve for the shape and BRDF is not just computationally expensive but also has a critical dependence on the ability to find a good initial solution since the underlying problem is non-convex and riddled with local minima. This naturally leads to a dense sampling of the 4-D BRDF space using precisely calibrated, and often prohibitively expensive, acquisition setups.

Uncalibrated lighting. Photometric stereo has also been addressed in the uncalibrated setting [50, 115, 145] where we do not have knowledge of the lighting directions. Alldrin et al. [11] formulate the problem as the minimization for the entropy and solve for the shape of the objects with Lambertian materials. For the non-Lambertian materials, Lu et al. [100] exploit the relation between surface normals and observed intensity profiles to estimate the shape of the object from multiple images. However, these approaches are inherently limited by bas-relief ambiguity, making the performance significantly worse (average 10 degree worse in angular errors as shown in [138]) than the calibrated counterpart.

BRDF acquisition using commodity devices. More recently, previous work has looked at the problem of reflectance capture "in the wild", under relatively unconstrained conditions, and using commodity hardware. Because of the ill-posed nature of this problem, these methods rely on extra information like the presence of reference materials in the scene [125] or restrict themselves to BRDFs with stochastic, texture-like spatial variations [7]. While their results are impressive, the use of reference materials or restrictive assumptions on the materials make these approaches less practical in real-world situations.

1.1.2 Illumination estimation

Real-world lighting often consists of multiple illuminants with different spectra. For example, outdoor illumination — both sunlight and skylight — differ in color temperature from indoor illuminants like incandescent, fluorescent, and LED lights. These variations in illuminant spectra manifest as color variations in captured images that are often a nuisance for vision-based analysis and photography. To this end, many techniques are devoted in the context of color constancy to remove the effect of the color of the light sources illuminating a scene.

Color constancy. While there are numerous techniques for performing white balancing, the vast majority of them assume that the captured scene is illuminated by a single light source [52, 55, 61]. Many real-world scenes are interesting precisely due to complex spatially-varying illumination with multiple light sources where the assumption of a single dominant light source is completely violated. This is commonly referred to as the "mixed illumination" scenario [27, 77, 126]. In the absence of additional assumptions or constraints, color constancy under mixed illumination is highly ill-posed and intractable since it is entirely possible that each scene point is illuminated by its own unique configuration of the light sources. As a consequence, the vast majority of prior techniques that perform white balancing method for mixed illumination rely either on user guidance [25, 27], or require knowledge of the number of light sources and their colors [77], or make simplifying assumptions that individual regions in the scene are illuminated by a single source [126].

Intrinsic images. More recently, intrinsic images techniques [18] have been proposed to directly separate an image into the reflectance and illumination layers. This is a particularly challenging decomposition because the effects of reflectance and illumination are combined into a single observation, which makes the inverse problem of separating them severely ill-posed. To tackle the problem, a common assumption made is that there is a single (usually white) illuminant and the materials for the objects are Lambertian [16, 18, 151]. Recently, more techniques introduce additional prior on depth, color variations or utilize deep neutral networks to directly regress the reflectance and illumination. While these techniques return high qual-

Illuminant	Method	Number of images	Output
Single	Gray-world	Single	Illumination
	Statistical prior [53]	Single	Illumination
	Intrinsic image [20]	Single	Reflectance& illumination
Multiple	Statistical prior [62]	Single	Illumination
	Known color [77]	Single	Illumination
	User-aided [26]	Single	Reflectance& illumination
	Flash/no-flash [80]	Two	Illumination
	Video-based [120]	Multiple	Illumination

Table 1.2: Illumination analysis methods for different ambient environments. Note that Automatic refers the methods which do not rely on extra prior knowledge of the scene while Prior refers the methods which require prior knowledge on the type or the color of the light sources. We also denote the user-aided methods by obtaining additional inputs from the users in the form of scribbles that mark regions with the same color or even regions that are known to be entirely white. For the flash/no-flash method, it requires a pair of flash/no-flash image for the same scene as the input. In comparison, video-based method focuses on the image sequence or a video recorded for a static scene. For our method, we aim for solving the problem by the use of a single image under the mixture of multiple light sources.

ity results, real world consists of objects with complex reflectance as well as being illuminated by multiple light sources with different colors, which significantly limits the applicability of the state-of-the-art techniques.

1.1.3 Post-capture editing of reflectance and illumination

One of the secondary objectives of this dissertation is to enable post-capture manipulation of a photograph and in particular the ability to change the reflectance and illumination in the scene. Current techniques for doing so rely heavily on user annotations. One approach is to obtain additional inputs from users in the form of scribbles that mark regions with similar reflectance or being illuminated with the same illuminant. User-guided approaches have been widely used in many image processing tasks including image editing [98], intrinsic images [25, 134], and color correction [26, 30]. In the context of white balancing, Boyadzhiev et al. [26] utilize user scribbles to indicate color attributes of the scene such as white surfaces and constant lighting regions. This enables an interpolation framework that propagates the information from user specified pixels to the under-determined regions. However, user guided approaches are not preferable for many reasons including the time and expertise required for a user to provide meaningful input. In many ways, we seek to automate this process by providing a robust framework for shape, reflectance and illumination estimation.

1.2 Contributions

We aim to solve for shape and reflectance via the use of easy-to-deploy capture devices, i.e. smart phones or tablets, as well as to extend the illumination analysis approaches by not only estimating the light colors but also separating the shadings induced by each light source. The key capabilities we seek to enable are in the form of editing the shape, reflectance and illumination in a scene. To this end, we make the following contributions in the dissertation.

1.2.1 Shape and spatially-varying reflectance estimation

We addresses the problem of estimating the shape of objects that exhibit spatially-varying reflectance.

Capture with photometric stereo setup. We assume that multiple images of the object are obtained under a fixed view-point and varying illumination, i.e., the setting of photometric stereo. At the core of our techniques is the assumption that the BRDF at each pixel lies in the non-negative span of a known BRDF dictionary. This assumption enables a per-pixel surface normal and BRDF estimation framework that is computationally tractable and requires no initialization in spite of the underlying problem being non-convex. Our estimation framework first solves for the surface normal at each pixel using a variant of example-based photometric stereo. We design an efficient multi-scale search strategy for estimating the surface normal and subsequently, refine this estimate using a gradient descent procedure. Given the surface normal estimate, we solve for the spatially-varying BRDF by constraining the BRDF at each pixel to be in the span of the BRDF dictionary; here, we use additional priors to further regularize the solution. A hallmark of our approach is that it does not require iterative optimization techniques nor the need for careful initialization, both of which are endemic to most state-of-the-art techniques.

Capture with mobile setup. We propose the use of a light-weight setup consisting of a collocated camera and light source commonly found on mobile devices to reconstruct surface normals and spatially-varying BRDFs of near-planar material samples. A collocated setup provides only a 1-D univariate sampling of a 3-D isotropic BRDF. We show that a univariate sampling is sufficient to estimate parameters of commonly used analytical BRDF models. Subsequently, we use a dictionary-based reflectance prior to derive a robust technique for per-pixel normal and BRDF estimation. We demonstrate real-world shape and capture, and its application to material editing and classification, using real data acquired using a mobile phone.

1.2.2 Illumination analysis

We address the problem of illumination analysis by estimating illuminant colors as well as analyzing and editing scene illuminants based on their spectral differences.

Estimate the illuminant colors. We present a technique to estimate the illumination colors for the images captured in the mixture of multiple light sources by leveraging flash photography. Even though this problem is severely ill-posed, we show that using two images — captured with and without flash lighting — leads to a closed form solution for spatially-varying mixed illumination. Our solution is completely automatic and makes no assumptions about the number or nature of the illuminants. We also propose an extension of our scheme to handle practical challenges such as shadows, specularities, as well as the camera and scene motion.

Separate the illuminant sources. We first leverage a flash/no-flash image pair to analyze and edit scene illuminants based on their spectral differences. We derive a novel physics-based relationship between color variations in the observed flash/no-flash intensities and the spectra and surface shading corresponding to individual scene illuminants. Our technique uses this constraint to automatically separate an image into constituent images lit by each illuminant. This separation can be used to support applications like white balancing, lighting editing, and RGB photometric stereo, where we demonstrate results that outperform state-of-the-art techniques on a wide range of images. We further extend the idea by using a single image to separate the illuminants. We do this by training a deep neural network to predict the per-pixel reflectance chromaticity of the scene, which we use in conjunction with a previous flash/no-flash image-based separation algorithm to produce the final two output images. We design our reflectance chromaticity network and loss functions by incorporating intuitions from the physics of image formation. We show that this leads to significantly better performance than other single image techniques and even approaches the quality of the two image separation method.

While ideas espoused in this dissertation provide important steps towards both reflectance, shape and illumination in the wild, our methods are still limited by particular assumptions we made and thus may not be well suited to especially complex scenes. To address these limitations, we may consider improving the current techniques in two ways. First, for the shape and reflectance estimation, instead of assuming single calibrated light source, we would like to extend the idea by considering the mixture of multiple light sources. To this end, we aim for a holistic approach that combines a large-scale training dataset, and incorporate the physically-motivated networks to efficiently handle this increased complexity. Second, as an extension for the lighting separation technique, we would like not only estimate the light color but also predict the position of the light source by incorporating the data-driven methods as shown in [59].

1.3 Organization

Chapter 2 studies the shape and SV-BRDF estimation via the photometric stereo setup. In Chapter 3, we extend the estimation to the mobile devices. Chapter 4 study the illuminant source separation for the scene under mixture of illumination via a flash/no-flash image pair. In Chapter 5, we enable such applicability for a single photograph by incorporating the deep neutral network. Finally, we include the discussions on the limitations of the proposed techniques and highlight future directions of research in Chapter 6 .

Chapter 2

Photometric stereo for spatially-varying BRDF

Photometric stereo [152] seeks to estimate the shape of an object from images obtained from a static camera and under varying lighting. While there has been remarkable progress in photometric stereo, the vast majority of techniques are devoted to scenes that exhibit simple reflectance properties. In particular, scenes with Lambertian reflectance have received the bulk of the attention [76, 149, 152] due to the immense simplification that such an assumption provides. Unfortunately, real-life scenes often involve non-Lambertian materials that interact with light in complex ways; this creates a significant disconnect between theory and practice.

In this chapter, we present a photometric stereo method for recovering the shape and the reflectance of opaque objects that exhibit spatially-varying reflectance. The key challenge here is that the reflectance, characterized in terms of spatially-varying bidirectional reflectance distribution function (SV-BRDF), and the shape, characterized in terms of surface normals, are inherently coupled and need to be estimated jointly. Further, the SV-BRDF is a 6D function of space and incident/outgoing angles and hence, can be very high-dimensional. In the absence of additional assumptions, estimating the SV-BRDF requires a large number of input images for robust estimation.

A common assumption for enabling computationally tractable models for SV-BRDF is that the BRDF at each pixel is a weighted combination of a *few, unknown reference BRDFs* [93]. The SV-BRDF is now represented using the reference BRDFs and their relative abundances at each pixel. This model offers a significant reduction in the dimensionality of the unknowns and, as a consequence, has been used in the context of photometric stereo [9, 63]. In Goldman et al. [63], the parametric isotropic Ward model [150] is used to characterize the reference BRDFs. Alldrin et al. [9] assume that the reference BRDFs are approximated by the non-parametric bivariate model [128] that approximates the 4D BRDF as a 2D signal. In both cases, the problem of shape and SV-BRDF estimation reduces to alternating minimization over the surface normals, the reference BRDFs, and abundances of the reference BRDFs at each pixel. The drawback of these approaches is that the optimization is not just computationally expensive but also has a critical dependence on the ability to find a good initial solution since the underlying problem is non-convex and riddled with local minima.

An alternate approach called example-based photometric stereo [71] introduces reference objects — typically, spherical objects — in the scene. This technique relies on the concept of *orientation consistency* [71] which suggests that two surface elements with identical normals and BRDFs will take the same appearance when placed in the same illumination. Example-based photometric stereo exploits orientation consistency as follows. Suppose that we want to estimate the surface normal at a particular pixel on the target. If the reference sphere has the same BRDF as the target, then we simply compare the intensity profile observed at each pixel on the sphere to that observed on the target pixel. The surface normal at the target pixel is recovered by finding the pixel on the sphere that best matches the intensity profile. In essence, each pixel on the reference sphere provides a candidate for the true surface normal. When the target's BRDF is spatially-varying, two reference objects — one diffuse and the one specular — can be used to recover the surface normals of the target by approximating the unknown BRDF at each pixel as a non-negative linear combination of the reference BRDFs [71]. A hallmark of example-based photometric stereo is that we do not need to calibrate the illumination. While example-based photometric stereo produces precise shape estimates without requiring the knowledge of lighting, there are multiple drawbacks associated with the method. The accuracy of recovering the surface normals is affected by the non-uniform sampling of normals of the spherical objects; specifically, we can expect to observe dense sampling of candidate normals along the viewing direction and coarse sampling near the vanishing directions. Many BRDFs are also poorly approximated as a linear combination of the two reference BRDFs. Finally, introducing reference objects is not always desirable in many practical applications.

The technique proposed in this chapter relies on the core principle of example-based photometric stereo *without actually introducing reference objects into the scene*. Given a dictionary whose atoms are BRDFs associated with a wide range of materials, we can render virtual spheres, one for each atom in the dictionary, under the knowledge of the scene illumination. This provides a set of "virtual exemplars" that can be used to obtain a per-pixel estimate of the shape and reflectance of the scene with arbitrary spatially-varying BRDF. The assumption that we make is that the unknown BRDF at each pixel lies in the non-negative span of the dictionary atoms. We show that the surface normals and the BRDFs can be estimated via a sequence of tractable linear inverse problems. This obviates the need for complex iterative optimization techniques as well as careful initialization required to avoid convergence to local minima. The interplay of these ideas for both the normal and SV-BRDF estimation provides not just a tractable solution to a previous ill-posed problem but also state-of-the-art results on challenging scenes (see Figure 2.1).

Contributions. We make the following contributions.

Model. We propose the use of a dictionary of BRDFs to regularize the surface normal and SV-BRDF estimation. The BRDF at each pixel of an object is assumed to lie in the non-negative span of the dictionary atoms.

Normal estimation. We show that the surface normal at each pixel can be efficiently estimated using a coarse-to-fine search and further refined using a gradient descent-based algorithm.

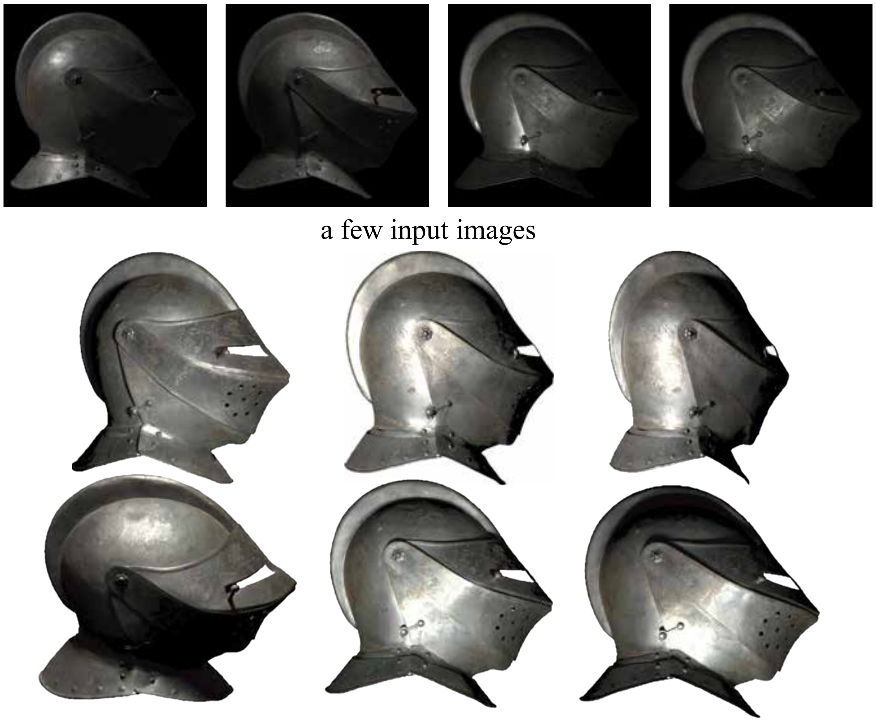
SV-BRDF estimation. Given the surface normals, we first recover the BRDF at each pixel independently by solving a linear inverse problem that enforces sparsity in the occurrence of the reference BRDFs at the pixel. To further regularize the BRDF estimation and obtain estimates with improved accuracy, we impose a low rank constraint on the SV-BRDF.

Validation. We showcase the accuracy of the shape and SV-BRDF estimation technique on a wide range of simulated and real scenes and demonstrate that the proposed technique outperforms state-of-the-art.

2.1 Prior work

In this section, we review some of the key techniques for shape estimation with respect to different BRDF models.

The diffuse + specular BRDF model. It is well known that the collection of images of a convex Lambertian object typically lies close to a low-dimensional subspace [19, 121]. This naturally leads to techniques [83, 155, 157] that robustly fit a low-dimensional subspace, capturing the Lambertian component while isolating non-Lambertian components, includ-



a few input images

object rendered in novel poses

Figure 2.1: We propose a framework for per-pixel estimation of surface normal and BRDF in the setting of photometric stereo. Shown above are the estimated shape and rendered images of a visually-complex object. The results were obtained from 250 images.

ing specularities, as sparse outliers. From the low-dimensional space reconstructed, they implement Lambertian photometric stereo to get the shape of objects. However, these techniques have restrictive assumptions on the range of BRDFs to which they are applicable, and more importantly, miss out on powerful cues to the shape of the object that are often present in specular highlights.

Parametric BRDF representations. Parametric models such as the

Blinn-Phong [22], Ward [150], Oren-Nayar [109], Ashikhmin-Shirley [13], Lafortune et al. [91], He et al. [70] and Cook-Torrance model [39] are based on macro-behavior established using specific micro-facet models on the materials, and have been widely used in computer graphics. In the context of shape and SV-BRDF estimation, Goldman et al. [63] utilize the isotropic Ward model [150] to reduce the dimensionality of the inverse problem. Oxholm and Nishino [111–113] further extend this idea by introducing a probabilistic formulation to estimate the BRDFs and exploit visual cues from multiple views under natural lighting conditions to reconstruct the object's shape. However, parametric models are inherently limited in their ability to provide precise approximations to the true BRDFs and further, often lead to challenging and ill-conditioned inverse problems.

Non-parametric BRDF representations. Non-parametric models are built upon the raw measured BRDFs [103, 107] and can provide faithful rendition to the empirical observations. The BRDFs are tabulated with respect to four angles, two for the incident direction and two for the outgoing direction. The high-dimensionality of non-parametric BRDF representations is often a challenge when we need to perform BRDF estimation, even when the shape is known.

Isotropic BRDFs. Isotropic materials exhibit a form of symmetry, wherein the reflectance of the material is unchanged when the incident and outgoing directions are jointly rotated about the surface normal. This enables the representation of isotropic BRDFs as the function over three as opposed to four angles. In the context of photometric stereo, Alldrin and Kriegman [10] observe that, for isotropic materials, the surface normal at each point can be restricted to lie on a plane. By restricting the light source with circular motion, Chandraker et al. [36] show that the shape can be estimated from the iso-contours of depth as well as an initial starting surface normal. When the isotropic BRDF has a single dominant lobe, Shi et al. [136] resolve the planar ambiguity and show that the surface normals can be uniquely determined. For the materials with multiple lobes, Shi et al. [137] address the problem by utilizing biquadratic to characterize the low-frequency components of isotropic materials, allowing for the normal estimation via solving a least square problem from the diffuse components. Ikehata and Aizawa [82] model the isotropic BRDFs as the sum of bivariate functions and solve for the surface normals via a constrained regression problem. Higo et al. [72] utilize properties of isotropy, visibility and monotonicity to restrict the solution space of the surface normal at each pixel. This enables a framework for shape estimation without the need for radiometric calibration. Lu et al. [99–101] further extend the idea by exploiting the relation between surface normals and observed intensity profiles to estimate the shape of the object from multiple images without illumination calibration. Finally, a bivariate approximation for isotropic materials is used in Romeiro et al. [128, 130] to estimate the BRDF of a known shape from a single image and without knowledge of the scene illumination.

Relationship to prior work. There have been other methods similar to our approach that seek to remove the use of "examples" from example-based photometric stereo. In Ackermann et al. [3,4], a partial reconstruction of the scene using multi-view stereo techniques is used as a reference (or example) to obtain dense normal estimates. In contrast, our technique focuses on the traditional problem of single-view photometric stereo. The assumption of the scene's reflectance function being composed of a few reference BRDFs is a common assumption used for photometric stereo under SV-BRDFs [9, 37, 63, 93, 161]. However, this leads to a multi-linear optimization in high-dimensional variables (the reference BRDFs) that is highly dependent on initial conditions. In contrast, our proposed technique avoids the need to estimate high-dimensional optimization by evoking knowledge of a dictionary of BRDFs.

2.2 Problem setup

Setup. We make the following assumptions, most of which are typical to photometric stereo-based shape estimation. First, the camera is orthographic and hence, the viewing direction $\mathbf{v} \in \mathbb{R}^3$ is constant across all scene points. Second, the scene illumination is assumed to be from a distant point light source. The light sources are assumed to be of constant brightness (equivalently, that calibration is known) and their direction is known. We denote $\mathbf{l}_k \in \mathbb{R}^3$ to refer to the lighting direction in the k-th image I^k. For

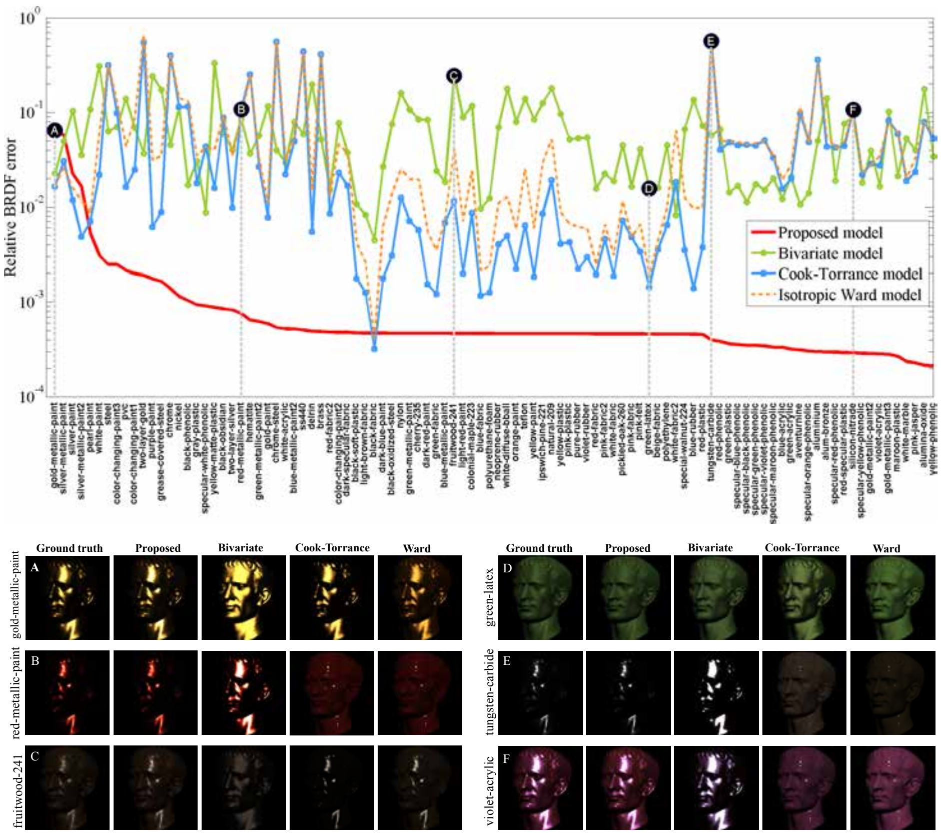


Figure 2.2: For the 100 materials in the database, we plot the approximation accuracy in relative RMS error [107] (also see (2.8)) for the proposed, bivariate [128], Cook-Torrance [39], and the isotropic Ward [150] models. For the proposed model, we use a leave-one-out scheme, wherein for each BRDF the remaining 99 BRDFs in the database are used to form the dictionary. The proposed model outperforms competing models both quantitatively (top) as well as in visual perception (bottom).

a light-stage, this information is typically obtained by a one-off calibration. Third, the effects of long-range illumination such as cast shadows and interreflections are assumed to be negligible; this is satisfied for objects with a convex shape. Finally, the radiometric response of the camera is linear.

BRDF representation. We follow the isotropic BRDF representation used in [133] in which a three-angle coordinate system based on half angles is used. Specifically, the BRDF is expressed as a function $\rho(\theta_h, \theta_d, \phi_d)$ with $\theta_h, \theta_d \in [0, \pi/2)$ and $\phi_d \in [0, 2\pi)$. However, by Helmholtz's reciprocity, the BRDF exhibits the following symmetry: $\rho(\theta_h, \theta_d, \phi_d) = \rho(\theta_h, \theta_d, \phi_d + \pi)$, and hence, it is sufficient to express $\phi_d \in [0, \pi)$. Following [103], we use a 1° sampling of each angle. As a consequence, a BRDF is represented as a point in a $T = 90 \times 90 \times 180 = 1,458,000$-dimensional space. When we deal with color images, we have a BRDF for each color channel and hence, the dimensionality of the BRDF goes up proportionally.

Consider a scene element with BRDF $\rho \in \mathbb{R}^T$, surface normal $\mathbf{n}$, illuminated by a point light source from a direction $\mathbf{l}$ and viewed from a direction $\mathbf{v}$. For this configuration of normal, incident light and viewing direction, the BRDF value is simply a linear functional of the vector ρ:

$$\mathbf{s}_{\{\mathbf{l},\mathbf{v};\mathbf{n}\}}^\top \rho,$$

where $\mathbf{s}_{\{\mathbf{l},\mathbf{v};\mathbf{n}\}}$ is a vector that encodes the geometry of the configuration. In essence, the vector samples the appropriate entry from ρ, allowing for the appropriate interpolation if the required value is off the sampling-grid.

Problem formulation. Our goal is to recover the surface normals and the SV-BRDF in the context of photometric stereo; i.e., multiple images of an object $\{I^1, \ldots, I^Q\}$ obtained from a static camera under varying lighting.

The intensity value $I_{\mathbf{p}}^{i}$ observed at pixel $\mathbf{p} = (x, y)$ with lighting $\mathbf{l}_i$ can be written as

$$I_{\mathbf{p}}^{i} = (\mathbf{s}_{\{\mathbf{l}_i,\mathbf{v};\mathbf{n}_{\mathbf{p}}\}}^{\top}\rho_{\mathbf{p}}) \cdot \max\{0, \mathbf{n}_{\mathbf{p}}^{\top}\mathbf{l}_i\}, \tag{2.1}$$

where $\rho_{\mathbf{p}}$ is the BRDF and $\mathbf{n}_{\mathbf{p}}$ is the surface normal at pixel $\mathbf{p}$, respectively, and $\max\{0, \mathbf{n}_{\mathbf{p}}^{\top}\mathbf{l}_i\}$ accounts for shading.

Given multiple intensity values at pixel $\mathbf{p}$, one for each lighting direction $\{\mathbf{l}_1, \ldots, \mathbf{l}_Q\}$, we can write

$$\begin{aligned} \mathbf{I}_{\mathbf{p}} &= \begin{pmatrix} I_{\mathbf{p}}^{1} \\ \vdots \\ I_{\mathbf{p}}^{Q} \end{pmatrix} = \begin{bmatrix} \max\{0, \mathbf{n}_{\mathbf{p}}^{\top}\mathbf{l}_1\} \cdot \mathbf{s}_{\{\mathbf{l}_1,\mathbf{v};\mathbf{n}_{\mathbf{p}}\}}^{\top} \\ \vdots \\ \max\{0, \mathbf{n}_{\mathbf{p}}^{\top}\mathbf{l}_Q\} \cdot \mathbf{s}_{\{\mathbf{l}_Q,\mathbf{v};\mathbf{n}_{\mathbf{p}}\}}^{\top} \end{bmatrix} \rho_{\mathbf{p}}, \\ &= A(\mathbf{n}_{\mathbf{p}})\rho_{\mathbf{p}}. \end{aligned} \tag{2.2}$$

Given the intensities, $\mathbf{I}_{\mathbf{p}}$, observed at a pixel $\mathbf{p}$ and knowledge of lighting directions $\{\mathbf{l}_1, \ldots, \mathbf{l}_Q\}$, we seek to estimate the surface normal $\mathbf{n}_{\mathbf{p}}$ and the BRDF $\rho_{\mathbf{p}}$ at the pixel. This problem is intractable without additional assumptions that constrain the BRDF to a lower-dimensional space.

Model for BRDF. The key assumption that we make is that the BRDF at a pixel $\mathbf{p}$ lies on the non-negative span of the atoms of a BRDF dictionary. Specifically, given dictionary $D = [\rho^1, \rho^2, \cdots, \rho^M]$, we assume that the BRDF at pixel $\mathbf{p}$ can be written as

$$\rho_{\mathbf{p}} = D\mathbf{c}_{\mathbf{p}}, \quad \mathbf{c}_{\mathbf{p}} \geq 0,$$

where $\mathbf{c_p} \in \mathbb{R}^M$ are the abundances of the dictionary atoms. In essence, we have constrained the BRDF to lie in an M-dimensional cone.[1] This provides immense reduction in the dimensionality of the unknowns at the expense of introducing a model misfit error. Indeed the success of this model relies on having a dictionary that is sufficiently rich to cover a wide range of interesting materials. Figure 2.2 shows the accuracy of various BRDF models on the MERL BRDF database [103].

In addition to the dictionary model for the BRDF, we also consider two additional priors.

- *Sparsity.* In the context of per-pixel BRDF estimation, we assume that $\mathbf{c_p}$ is sparse, suggesting that BRDF at the pixel $\mathbf{p}$ is the linear combination of a *few* dictionary atoms. The sparsity constraint avoids over-fitting to the intensity measurements $\mathbf{I_p}$ as well as provides a regularization for under-determined problems.

- *Low rank.* In the context of estimating the SV-BRDF for all pixels jointly, we assume that coefficient matrix $\mathbf{C} = [\mathbf{c_{p_1}}, \mathbf{c_{p_2}}, \ldots \mathbf{c_{p_N}}]$, that denotes the collection of the abundances for all the N pixels in the scene, is low rank. The low-rank prior on $\mathbf{C}$ implies that BRDFs at all pixels can be expressed as a linear combination of small number of unique reflectance functions. This prior is at the heart of many approaches for photometric stereo under

[1] A more appropriate model for the BRDF is that $(D\mathbf{c}) \geq 0$. However, this leads to significantly higher-dimensional constraints. We instead use a sufficient condition to achieve this, $\mathbf{c} \geq 0$.

SV-BRDF [9,63,93,161]. The low-rank prior also enables us to efficiently pool together information from multiple pixels, thereby providing significant improvements over the per-pixel estimates, without exploiting any explicit spatial smoothness priors.

Solution outline. We formulate the per-pixel surface normal and BRDF estimation using the following optimization problem.

$$\begin{aligned} \{\widehat{\mathbf{n}}_{\mathbf{p}}, \widehat{\mathbf{c}}_{\mathbf{p}}\} = \arg\min_{\mathbf{n},\mathbf{c}} \|\mathbf{I}_{\mathbf{p}} - A(\mathbf{n})D\mathbf{c}\|_2^2 + \lambda\|\mathbf{c}\|_1 \\ \text{s.t} \quad \mathbf{c} \geq 0, \|\mathbf{n}\|_2 = 1. \end{aligned} \tag{2.3}$$

The ℓ_1-penalty serves to enforce sparse solutions, with $\lambda \geq 0$ determining the level of sparsity in the solution. The optimization problem in (2.3) is non-convex due to unit-norm constraint on the surface normal $\mathbf{n}$ as well as the term $A(\mathbf{n})D\mathbf{c}$. Our solution methodology consists of two steps:

1. *Surface normal estimation.* We perform an efficient multi-scale search together with the gradient descent based refinement scheme which provides us with a precise estimate of the surface normal at pixel $\mathbf{p}$ (see Section 2.3.3);

2. *BRDF estimation.* We first solve (2.3) only over $\mathbf{c}$ with the normal fixed to obtain the BRDF at $\mathbf{p}$. We next incorporate a low rank constraint on the SV-BRDF to further regularize the BRDF estimates (see Section 2.4.2).

2.3 Surface normal estimation

In this section, we describe an efficient per-pixel surface normal estimation algorithm.

2.3.1 Virtual exemplar-based normal estimation

The first step of our surface normal estimation can be viewed as an extension of the method proposed in [71], where two spheres — one diffuse and one specular — are introduced in a scene along with the target object. Recall that, the scene is observed under Q different illuminations. Hence, at a pixel $\mathbf{p}$ on the target, we can construct the intensity profile $\mathbf{I}_\mathbf{p} \in \mathbb{R}^Q$ that enumerates the Q intensity values observed at $\mathbf{p}$. To obtain the surface normals at the pixel $\mathbf{p}$, we compare the intensity profile, $\mathbf{I}_\mathbf{p}$, to those on the reference spheres. The reference spheres provide a sampling of the space of the normals and hence, we can simply treat them as a collection of candidate normals $\mathcal{N}$. By orientation consistency, the surface normal estimation now reduces to finding the candidate normal that can best explain the intensity profile $\mathbf{I}_\mathbf{p}$. Given a candidate normal $\widetilde{\mathbf{n}}$, we have two intensity profiles, $\mathbf{I}_D(\widetilde{\mathbf{n}})$ and $\mathbf{I}_S(\widetilde{\mathbf{n}})$, one each for the diffuse and specular sphere, respectively. The estimate of the surface normal at pixel $\mathbf{p}$ is given as

$$\widehat{\mathbf{n}}_\mathbf{p} = \arg\min_{\widetilde{\mathbf{n}} \in \mathcal{N}} \min_{a_1, a_2 \geq 0} \|\mathbf{I}_\mathbf{p} - a_1 \mathbf{I}_D(\widetilde{\mathbf{n}}) - a_2 \mathbf{I}_S(\widetilde{\mathbf{n}})\|.$$

In [71], this is solved by scanning over all the pixels/candidate normals on the reference spheres.

Rendering virtual spheres. We rely on the same approach as [71] with the key difference that we virtually render the reference spheres. The virtual spheres are rendered as follows. Given the lighting directions $\{\mathbf{l}_1, \ldots, \mathbf{l}_Q\}$ and the BRDF dictionary $D = [\rho^1, \ldots, \rho^M]$, for each candidate normal $\widetilde{\mathbf{n}} \in \mathcal{N}$, we render a matrix $B(\widetilde{\mathbf{n}}) = [b_{ij}(\widetilde{\mathbf{n}})] \in \mathbb{R}^{Q \times M}$ such that $b_{ij}(\widetilde{\mathbf{n}})$ is the intensity observed at a surface with normal $\widetilde{\mathbf{n}}$ and BRDF ρ^j, under lighting $\mathbf{l}_i$.

$$b_{ij}(\widetilde{\mathbf{n}}) = \max\{0, \widetilde{\mathbf{n}}^\top \mathbf{l}_i\} \cdot \mathbf{s}^\top_{\{\mathbf{l}_i, \mathbf{v}; \widetilde{\mathbf{n}}\}} \rho^j,$$

We render one such matrix $B(\cdot)$ for each candidate normal in $\mathcal{N}$. Given these virtually rendered spheres, we can solve (2.3) by searching over all candidate normals.

Brute-force search. For computationally efficiency, we drop the sparsity-promoting term in (2.3). We empirically observed that this makes little difference in the estimated surface normals. Now, given the intensity profile $\mathbf{I_p}$ at pixel $\mathbf{p}$ and noting that $B(\widetilde{\mathbf{n}}) = A(\widetilde{\mathbf{n}})D$, solving (2.3) reduces to:

$$\widehat{\mathbf{n}}_\mathbf{p} = \underset{\widetilde{\mathbf{n}} \in \mathcal{N}}{\arg\min} \quad \min_{\mathbf{c} \geq 0} \quad \|\mathbf{I_p} - B(\widetilde{\mathbf{n}})\mathbf{c}\|_2^2. \tag{2.4}$$

The unit-norm constraint on the surface normals is absorbed into the candidate normals being unit-norm. The optimization problem in (2.4) requires solving a set of non-negative least squares (NNLS) sub-problems, one for each element of $\mathcal{N}$. For the results in the chapter, we used the `lsnonneg` function in MATLAB to solve the NNLS sub-problems.

The accuracy and the computational cost in solving (2.4) depends solely on the cardinality of the candidate set $\mathcal{N}$, $|\mathcal{N}|$. We obtain $\mathcal{N}$ by uniform or equi-angular sampling on the sphere [67]. Note that the smaller the angular spacing of $\mathcal{N}$, the larger is its cardinality. For example, a 5° equi-angular sampling over the hemisphere requires approximately 250 candidates, while a 0.5° requires $20,000$ candidates. Given that the time-complexity of the brute-force search is linear in $|\mathcal{N}|$, the computational costs for obtaining very precise normal estimates can be overwhelming (see Table 2.1). To alleviate this, we outline a coarse-to-fine search strategy that is remarkably faster than the brute-force approach with little loss in accuracy.

2.3.2 Coarse-to-fine search

Figure 2.3 shows the value of

$$E(\widetilde{\mathbf{n}}) = \min_{\mathbf{c} \geq 0} \|\mathbf{I_p} - B(\widetilde{\mathbf{n}})\mathbf{c}\|$$

as a function of the candidate normal $\widetilde{\mathbf{n}}$ for a few examples. We observe that there is a gradual increase in error value as we moved away from the global minima of $E(\cdot)$. We exploit this to design a coarse-to-fine search strategy where we first evaluate the candidate normals at a coarse sampling and subsequently search in the vicinity of this solution at a finer sampling.

Specifically, let $\mathcal{N}_\theta$ be the set of equi-angular sampling on the unit-sphere where the angular spacing is θ degrees. Given a candidate normal $\widetilde{\mathbf{n}}$, we define

$$C_\theta(\widetilde{\mathbf{n}}) = \{\mathbf{n} \mid \langle \mathbf{n}, \widetilde{\mathbf{n}} \rangle \geq \cos\theta,\ \|\mathbf{n}\|_2 = 1\}$$

as the set of unit-norm vectors within θ-degrees from $\widetilde{\mathbf{n}}$,

In the first iteration, we initialize the candidate normal set $\mathcal{N}^{(1)} = \mathcal{N}_{\theta_1}$. Now, at the j-th iteration, we solve (2.4) over a candidate set $\mathcal{N}^{(j)}$. Suppose that $\widehat{\mathbf{n}}^{(j)}$ is the candidate normal where the minimum occurs at the j-th iteration. The candidate set for the $(j+1)$-th iteration is constructed as

$$j \geq 1, \quad \mathcal{N}^{(j+1)} = C_{\theta_j}(\widehat{\mathbf{n}}^{(j)}) \cap \mathcal{N}_{\theta_{j+1}},$$

with $\theta_{j+1} < \theta_j$. That is, the candidate set is simply the set of all candidates at a finer angular sampling that are no greater than the current angular sampling from the current estimate. This is repeated till we reach the finest resolution at which we have candidate normals. For the results in this chapter, we use the following values: $\theta_1 = 10°, \theta_2 = 5°, \theta_3 = 3°, \theta_4 = 1°$, and $\theta_5 = 0.5°$. For efficient implementation, we pre-render $B(\widetilde{\mathbf{n}})$ for $\widetilde{\mathbf{n}} \in \mathcal{N}_{\theta_1} \cup \cdots \cup \mathcal{N}_{\theta_5}$.

The computational gains obtained via this coarse-to-fine search strategy are immense. Table 2.1 shows the run-time and precision of both brute force and coarse-to-fine normal estimation strategy for different levels of angular sampling in the generation of the candidate normal set. As expected the

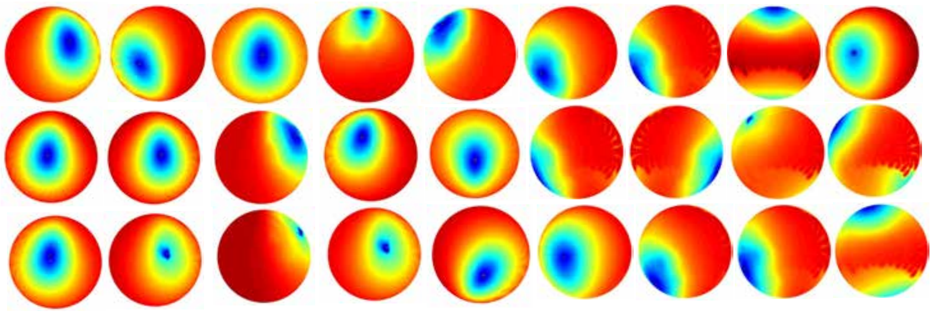


Figure 2.3: We can observe that the global minima is compact and the error increases largely monotonically in its vicinity. This motivates our coarse-to-fine search strategy.

		θ_1	θ_2	θ_3	θ_4	θ_5
Brute force	time	0.18s	0.77s	4.25s	27.3s	74.1s
	ang. error	7.07°	3.99°	1.56°	0.60°	0.42°
	max samples	76	327	1828	11829	31830
Coarse to fine	time	0.18s	0.19s	0.23s	0.34s	0.41s
	ang. error	7.07°	4.99°	2.56°	1.23°	0.82°
	max samples	76	81	89	105	112

Table 2.1: Comparison of brute-force and coarse-to-fine normal estimation for different angular samplings of the candidate normals: $\theta_1 = 10^\circ, \theta_2 = 5^\circ, \theta_3 = 3^\circ, \theta_4 = 1^\circ$, and $\theta_5 = 0.5^\circ$. For each method, we report the time taken, the angular error, and the maximum number of candidates evaluated. Shown are averages over 100 random trials.

run-time of the brute force algorithm is linear in the number of candidates. In contrast, the coarse-to-fine strategy requires a tiny fraction of this time while nearly achieving the same precision as the brute force strategy.

A drawback of both the brute-force as well as the coarse-to-fine approaches is that the estimated normals are restricted by the candidate normal set and hence, the accuracy of the estimates, on an average, cannot be better than the half the angular spacing of the candidate set at the finest level. To address this, we propose a local descent-based scheme that circumvents the limitations of using just candidate normals.

2.3.3 Gradient descent-based normal estimation

Our gradient descent-based scheme to estimate the surface normals relies on two observations: first, we can use the estimate obtained from the coarse-to-fine strategy as an accurate initial guess; and second, we can linearize the cost function in (2.4) in the vicinity of our initial guess and devise a gradient descent algorithm.

Specifically, let $f(\mathbf{n_p}, \mathbf{c_p})$ be the value of the data term in (2.4), i.e.

$$f(\mathbf{n_p}, \mathbf{c_p}) = \|\mathbf{I_p} - B(\mathbf{n_p})\mathbf{c_p}\|_2^2.$$

Now, at $\widehat{\mathbf{n}}_\mathbf{p}, \widehat{\mathbf{c}}_\mathbf{p}$ obtained by using coarse-to-fine search, we can linearize $f(\mathbf{n_p}, \mathbf{c_p})$ as

$$f(\widehat{\mathbf{n}}_\mathbf{p} + \triangle\mathbf{n_p}, \widehat{\mathbf{c}}_\mathbf{p} + \triangle\mathbf{c_p}) = \|\mathbf{I_p} - B(\widehat{\mathbf{n}}_\mathbf{p} + \triangle\mathbf{n_p})(\widehat{\mathbf{c}}_\mathbf{p} + \triangle\mathbf{c_p})\|_2^2.$$

Given $B(\widehat{\mathbf{n}}_\mathbf{p})$ is locally smooth[2], it can be linearized at $\widehat{\mathbf{n}}_\mathbf{p}$ as

$$B(\widehat{\mathbf{n}}_\mathbf{p} + \triangle\mathbf{n}_\mathbf{p}) = B(\widehat{\mathbf{n}}_\mathbf{p}) + \nabla_\mathbf{n} B(\widehat{\mathbf{n}}_\mathbf{p})\triangle\mathbf{n}_\mathbf{p}.$$

To account for the unit norm constraint on $\widehat{\mathbf{n}}_\mathbf{p} + \triangle\mathbf{n}_\mathbf{p}$, we utilize the elevation angle, which is denoted as θ, and the azimuth angle, which is denoted as ϕ, to represent surface normals. That is, we restrict the update of surface normals into a two dimensional space by absorbing the unit norm constraint. In particular, we can write $B(\widehat{\mathbf{n}}_\mathbf{p} + \triangle\mathbf{n}_\mathbf{p})$ as

$$B(\widehat{\mathbf{n}}_\mathbf{p} + \triangle\mathbf{n}_\mathbf{p}) = B(\widehat{\mathbf{n}}_\mathbf{p}) + \nabla_\phi B(\widehat{\mathbf{n}}_\mathbf{p})\triangle\phi_\mathbf{p} + \nabla_\theta B(\widehat{\mathbf{n}}_\mathbf{p})\triangle\theta_\mathbf{p},$$

where $\triangle\phi_\mathbf{p}$ and $\triangle\theta_\mathbf{p}$ denote local gradients for the elevation and azimuth angles of $\widehat{\mathbf{n}}_\mathbf{p}$, respectively. In essence, we have now reformulated the problem in (2.4) into a form involving the local gradients in surface normals and abundances. This enables us to refine the normal estimates without any restrictions imposed by the sampling of the candidate set.

Now, an estimate of local gradients at a pixel $\mathbf{p}$ can be obtained by solving

$$\begin{aligned}
&\{\triangle\widehat{\theta}_\mathbf{p}, \triangle\widehat{\phi}_\mathbf{p}, \triangle\widehat{\mathbf{c}}_\mathbf{p}\} = \underset{\triangle\theta,\triangle\phi,\triangle c}{\arg\min} \\
&\|\mathbf{I}_\mathbf{p} - (B(\widehat{\mathbf{n}}_\mathbf{p}) + \nabla_\phi B(\widehat{\mathbf{n}}_\mathbf{p})\triangle\phi + \nabla_\theta B(\widehat{\mathbf{n}}_\mathbf{p})\triangle\theta)(\widehat{\mathbf{c}}_\mathbf{p} + \triangle\mathbf{c})\|_2^2 \qquad (2.5) \\
&\text{s.t} \quad \widehat{\mathbf{c}}_\mathbf{p} + \triangle\mathbf{c} \geq 0.
\end{aligned}$$

[2] Although $B(\mathbf{n})$ involves the shading term, the smoothness property holds in most scenarios due to the dense angular sampling of the candidate normals.

We drop the second-order terms $\triangle\theta\triangle\mathbf{c}$ and $\triangle\phi\triangle\mathbf{c}$ in (2.5), which contributes little energy to the cost function, and we can solve the resulting convex optimization problem in (2.5) over $\triangle\theta$, $\triangle\phi$ and $\triangle\mathbf{c}$ using alternating minimization.

Estimating $\nabla_\phi B(\widehat{\mathbf{n}}_\mathbf{p})$ and $\nabla_\theta B(\widehat{\mathbf{n}}_\mathbf{p})$. Let $\mathcal{N}^{(J)}$ be the candidate normals set at the finest sampling level J. To estimate the gradients at the current estimate $\widehat{\mathbf{n}}_\mathbf{p}$, we construct a set $\mathcal{S} \subset \mathcal{N}^{(J)}$ of all normals in $\mathcal{N}^{(J)}$ that lie in a small neighborhood (smaller than 2 degrees in angular difference) of $\widehat{\mathbf{n}}_\mathbf{p}$. For each normal $\widetilde{\mathbf{n}} \in \mathcal{S}$ we can write

$$B(\widetilde{\mathbf{n}}) - B(\widehat{\mathbf{n}}_\mathbf{p}) = (\widetilde{\phi} - \widehat{\phi}_\mathbf{p})\nabla_\phi B(\widehat{\mathbf{n}}_\mathbf{p}) + (\widetilde{\theta} - \widehat{\theta}_\mathbf{p})\nabla_\theta B(\widehat{\mathbf{n}}_\mathbf{p}),$$

where $(\widetilde{\theta}, \widetilde{\phi})$ is the Euler angle representation of $\widetilde{\mathbf{n}}$. We can set up an overdetermined set of equations by stacking together the constraints arising from normals in the set $\mathcal{S}$ to estimate the gradients, $\nabla_\phi B(\widehat{\mathbf{n}}_\mathbf{p})$ and $\nabla_\theta B(\widehat{\mathbf{n}}_\mathbf{p})$. We recover the gradients by taking the pseudo-inverse of this overdetermined linear system.

Given the estimated $\nabla_\phi B(\widehat{\mathbf{n}}_\mathbf{p})$ and $\nabla_\theta B(\widehat{\mathbf{n}}_\mathbf{p})$, we perform the following steps until convergence.

Updating $\triangle\widehat{\phi}_\mathbf{p}$ and $\triangle\widehat{\theta}_\mathbf{p}$. Both $\triangle\widehat{\phi}$ and $\triangle\widehat{\theta}$ are estimated by solving a least square problem.

Update $\triangle\widehat{\mathbf{c}}_\mathbf{p}$. Due to the non-negative constraint on $\widehat{\mathbf{c}}_\mathbf{p} + \triangle\mathbf{c}$, we first solve the least square problem over $\triangle \mathbf{c}$ and then project the solution to the space specified by the constraint.

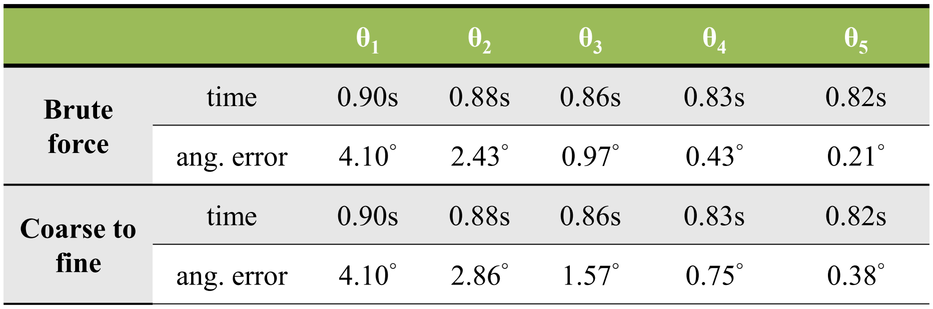
The estimate of surface normals can be obtained by

$$\widehat{\theta}_{\mathbf{p}} \leftarrow \widehat{\theta}_{\mathbf{p}} + \triangle\widehat{\theta}_{\mathbf{p}},$$

$$\widehat{\phi}_{\mathbf{p}} \leftarrow \widehat{\phi}_{\mathbf{p}} + \triangle\widehat{\phi}_{\mathbf{p}},$$

$$\widehat{\mathbf{n}}_{\mathbf{p}} \leftarrow [\cos(\widehat{\phi}_{\mathbf{p}})\sin(\widehat{\theta}_{\mathbf{p}}), \sin(\widehat{\phi}_{\mathbf{p}})\sin(\widehat{\theta}_{\mathbf{p}}), \cos(\widehat{\theta}_{\mathbf{p}})]^{\top}.$$

Observations. The gradient descent procedure described above can be solved efficiently. For a single surface normal, optimization to converge takes between 0.8 and 0.9 seconds in MATLAB on a desktop with Intel Xeon 3.6G CPU. In Table 2.2, we tabulate the improvements provided by the gradient descent procedure when initialized with the solutions of the brute force as well as the coarse-to-fine strategies. We observe that both algorithms benefit immensely from utilizing the gradient descent search. Further, the average error can be made smaller than the sampling resolution of candidate normals on the unit sphere. However, the average angular error often does not reduce to zero due to measurement noise as well as model misfit.



		θ_1	θ_2	θ_3	θ_4	θ_5
Brute force	time	0.90s	0.88s	0.86s	0.83s	0.82s
	ang. error	4.10°	2.43°	0.97°	0.43°	0.21°
Coarse to fine	time	0.90s	0.88s	0.86s	0.83s	0.82s
	ang. error	4.10°	2.86°	1.57°	0.75°	0.38°

Table 2.2: Gradient descent local search starting from both the brute-force and coarse-to-fine normal estimation for different angular samplings in the candidate normals: $\theta_1 = 10°, \theta_2 = 5°, \theta_3 = 3°, \theta_4 = 1°$, and $\theta_5 = 0.5°$. Shown are aggregate statistics over 100 randomly generated trials.

2.4 Reflectance estimation

2.4.1 Per-pixel BRDF estimate

Given the surface normal estimate $\widehat{\mathbf{n}}_{\mathbf{p}}$, we obtain an estimate of the BRDF at each pixel, individually, by solving

$$\widehat{\mathbf{c}}_{\mathbf{p}} \;= \arg\min_{\mathbf{c}\geq 0} \|\mathbf{I}_{\mathbf{p}} - B(\widehat{\mathbf{n}}_{\mathbf{p}})\mathbf{c}\|_2^2 + \lambda\|\mathbf{c}\|_1. \tag{2.6}$$

The use of the ℓ_1-regularizer promotes sparse solutions and primarily helps in avoiding over-fitting to the observed intensities. The optimization problem in (2.6) is convex and we used CVX [65], a general purpose convex solver, to obtain solutions. The estimate of the BRDF at pixel $\mathbf{p}$ is given as $\widehat{\rho}_{\mathbf{p}} = D\widehat{\mathbf{c}}_{\mathbf{p}}$. The value of λ was manually tuned for best performance on synthetic data. For color-imagery, we solve for the coefficients associated with each color channel separately.

The advantage of the per-pixel BRDF estimation framework is the ability to handle arbitrarily complex spatial variations in the BRDF. However, a drawback of per-pixel BRDF estimation is the relative lack of information available at a pixel. When we know a priori that multiple pixels share the same BRDF, then we can solve (2.6) simply by concatenating their corresponding intensity profiles and their respective $B(\cdot)$ matrices. As is to be expected, pooling intensities observed at multiple pixels significantly improves the quality of the estimates. Yet, while spatial averaging or spatial smoothness priors improve the quality of the estimate, inherently they require the object to exhibit smooth spatial-variations in its BRDF. To address this, we pool together information across multiple pixels by utilizing the low rank prior.

2.4.2 Incorporating low rank priors

Given the matrix $\mathbf{C} = [\mathbf{c}_{\mathbf{p}_1}, \mathbf{c}_{\mathbf{p}_2}, \ldots, \mathbf{c}_{\mathbf{p}_N}]$ for the estimated abundances for all N pixels in the scene, we constrain $\mathbf{C}$ to be low rank. The low rank prior, inspired by prior work [9, 63, 93, 146], suggests that the SV-BRDF of the scene under consideration can be generated from a small number of unique reference reflectance functions such that linear combinations of these reference BRDFs produces the BRDF at any pixel. The low rank prior also allows us to restrict the solution space for all the pixels globally without enforcing any spatial smoothness or clustering of pixels.

Solution outline. We can now formulate a global optimization problem that incorporates the low rank prior as follows: The estimate of the abundances of the BRDF at pixel $\mathbf{p}$ is given as

$$\begin{aligned}\widehat{\mathbf{C}} &= \arg\min_{\mathbf{C}} \beta \|\mathbf{C}\|_* + \sum_{\mathbf{p}} \|\mathbf{I}_{\mathbf{p}} - B(\widehat{\mathbf{n}}_{\mathbf{p}})\mathbf{c}_{\mathbf{p}}\|_2^2 + \lambda \|\mathbf{c}_{\mathbf{p}}\|_1. \\ &\qquad\qquad \text{s.t.} \qquad \mathbf{C} = [\mathbf{c}_{\mathbf{p}_1}, \ldots, \mathbf{c}_{\mathbf{p}_N}], \quad \forall \mathbf{p}, \mathbf{c}_{\mathbf{p}} \geq 0. \end{aligned} \tag{2.7}$$

where $\|\mathbf{C}\|_*$, the nuclear norm of the matrix $\mathbf{C}$, promotes low-rank solutions [29, 51, 123]. Note that we do not control the rank of the solution directly, but instead do so by using the penalty parameter β. This is achieved as follows: for large values of β, the nuclear norm penalty is strongly enforced and hence, we can expect the solution to be of low-rank. Similarly, small values of β lead to solutions with larger rank. We exploit this observation as a precept in sequentially selecting β till we find a solution of desired rank.

The objective in (2.7) consists of a smooth data fidelity term and two non-differentiable regularization terms, the ℓ_1-term that promotes sparse solutions and the nuclear norm that promotes low-rank solutions. We solve this by using prox-linear or forward-backward operator splitting [116]. This results in the following algorithm.

Given the estimate in the (k)-th iteration, $\widehat{\mathbf{C}}^{(k)} = [\widehat{\mathbf{c}}_{\mathbf{p}_1}^{(k)}, \ldots, \widehat{\mathbf{c}}_{\mathbf{p}_N}^{(k)}]$, we perform three operations to obtain the estimate at the $(k+1)$-th iteration.

- **Gradient descent.** We perform the "forward operation" which comprises of a gradient descent on the smooth data fidelity term. Given the separable nature of the data fidelity term, we can apply this on each pixel separately.

$$\widehat{\mathbf{a}}_{\mathbf{p}}^{(k+1)} = \widehat{\mathbf{c}}_{\mathbf{p}}^{(k)} + 2tB(\widehat{\mathbf{n}}_{\mathbf{p}})^\top(\mathbf{I}_{\mathbf{p}} - B(\widehat{\mathbf{n}}_{\mathbf{p}})\widehat{\mathbf{c}}_{\mathbf{p}}^{(k)}),$$

where t denotes the update step in gradient descent.

- **Soft thresholding.** We next perform the first "backward operation" corresponding to the ℓ_1-norm and the non-negativity constraint. The associated proximal operator results in soft thresholding at each pixel, followed by a thresholding at zero to enforce non-negativity

  $$\widehat{\mathbf{b}}_{\mathbf{p}}^{(k+1)} = \max\left(\mathcal{S}_{\lambda t}\left(\widehat{\mathbf{a}}_{\mathbf{p}}^{(k+1)}\right), \mathbf{0}\right),$$

  where $\mathcal{S}_\tau(\cdot)$ denotes the soft-thresholding operator defined as $\mathcal{S}_\tau(\mathbf{x}) = \operatorname{sgn}(\mathbf{x})\max(|\mathbf{x}| - \tau, 0)$, and $|\cdot|$ is the absolute value.

- **Singular value thresholding.** Finally, we perform the second "backward operation" corresponding to the nuclear norm. The associated proximal operator results in a singular value thresholding step. Let $\widehat{\mathbf{B}}^{(k+1)} = [\widehat{\mathbf{b}}_{\mathbf{p}_1}^{(k+1)}, \ldots, \widehat{\mathbf{b}}_{\mathbf{p}_N}^{(k+1)}]$. Now, we can obtain $\widehat{\mathbf{C}}^{(k+1)}$ as

  $$\widehat{\mathbf{C}}^{(k+1)} = U[\mathcal{S}_\beta(\sigma)]V^\top,$$

  where $\widehat{\mathbf{B}}^{(k+1)} = U\text{diag}(\sigma)V^\top$.

We perform the update for the matrix **C** until the convergence can be reached. In the next section, we carefully characterize the performance of our proposition using synthetic and real examples.

2.5 Results

We characterize the performance of our technique using both synthetic and real datasets.

2.5.1 Synthetic experiments

We use the BRDFs in the MERL database [103] in a leave-one-out scheme for testing the accuracy of our proposed algorithms. Specifically, when we simulate a test object using a particular material, the dictionary is comprised of BRDFs of the remaining $M = 99$ materials from the database. We used the configuration in the light-stage described in [47] for our collection of lighting directions.

Performance of normal estimation

We characterize the performance of normals estimation by testing on the synthetic data with varying number of lighting directions, varying BRDFs as well as varying dictionary size and type.

Varying number of input images. Figure 2.4 characterizes the errors in surface normal for varying number of input images or equivalently, lighting

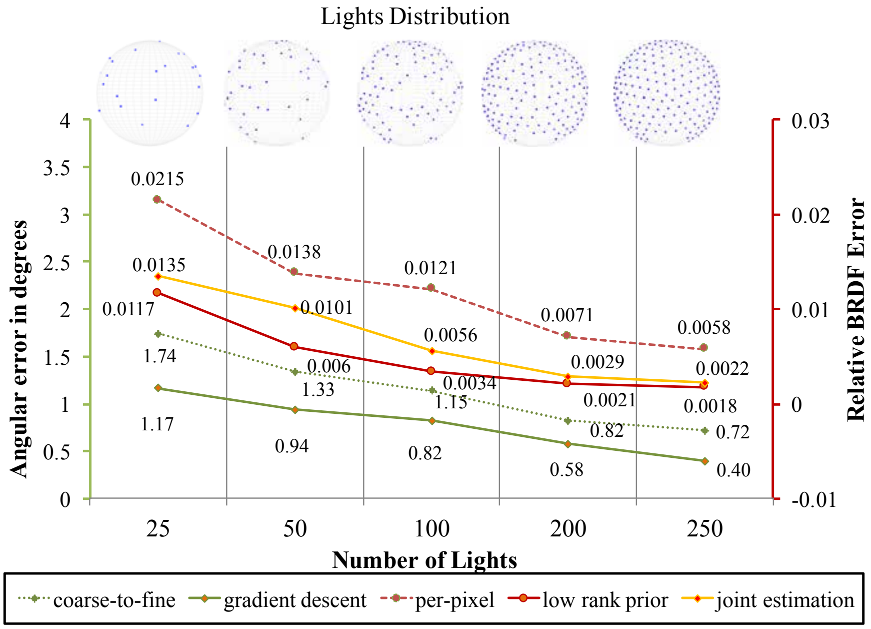


Figure 2.4: We estimate the angular errors for the coarse-to-fine (in dot green) and the gradient descent method (solid green line). We also estimate the relative BRDF errors for both per-pixel (in dot red) and rank-1 prior (solid red line) under perfect knowledge of the surface normals. Finally, we test the entire estimation pipeline by measuring the accuracy of BRDFs using the surface normals from the gradient descent scheme (orange solid line). The plots were obtained by averaging across all 100 BRDFs in the MERL database and 20, 000 randomly-generated normals per material.

directions. We report the average angular error for both the coarse-to-fine search strategy and gradient descent method. In each case, the average angular error is computed by randomly generating 20, 000 normals per material and varying across all 100 material BRDFs in the database. This experiment is similar in setup to the one reported in [136] which, to our knowledge, is one of the most accurate techniques for photometric stereo on isotropic BRDFs.

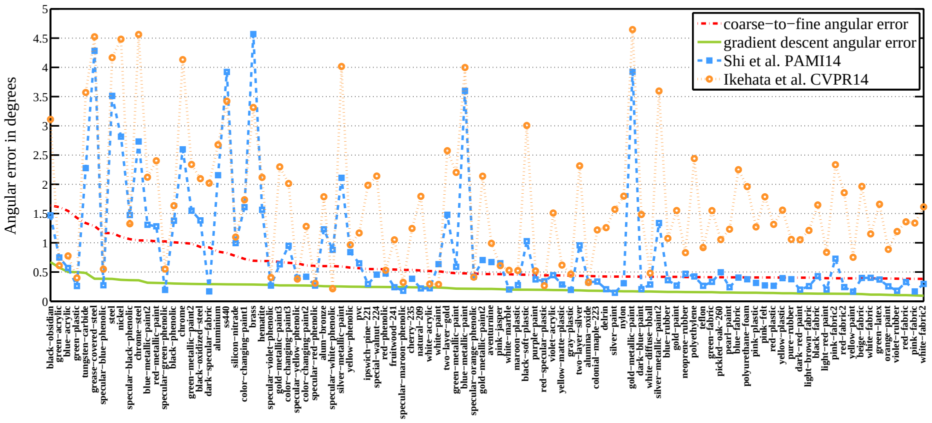


Figure 2.5: We fix the number of input images/lighting directions to 253. For each material BRDF, we compute average error over 1,000 randomly-generated surface normals for both the coarse-to-fine search strategy (in red) and the gradient descent method (in green). The gradient descent scheme outperforms both competing methods [82, 137] in 88 out of 100 materials.

In [136], for 200 images, the angular error in estimating only the elevation angle *when the azimuth is known* is reported as 0.88°; in contrast, our proposed technique, *without any prior knowledge of the azimuth*, has an angular error of 0.82° for the coarse-to-fine search and 0.58° for the gradient descent refinement.

Varying BRDF. Figure 2.5 compares the performance of the proposed technique to that of state-of-the art non-Lambertian photometric stereo algorithms [82, 137]. We fixed the number of images at $Q = 253$. Shown are aggregate statistics computed over 1,000 randomly-generated surface normals. For the coarse-to-fine technique, the worst-case error is less than 2° and, further, the error tapers down to 0.5° — which is the finest sampling of the candidate normals. Incorporating the gradient descent method provides substantial improvements and the angular error, in the best-case scenario, is reduced to 0.1°, which is much smaller than the finest sampling on the candidate normals. This demonstrates the value of the gradient descent method over only coarse-to-fine search strategy. We also note that the proposed technique algorithm outperforms the both state-of-the-art techniques [82,137] for most of materials we compare against; in total, the proposed technique has worse performance in 12 out of 100 materials. In addition to these simulations, in the supplemental material, we report the performance of the proposed normal estimation techniques as well as competing algorithms for multiple non-Lambertian BRDFs.

Varying dictionary size and type. Figure 2.6 evaluates the performance of both coarse-to-fine and gradient descent approaches as the number of dictionary atoms is varied. We use the same setting as Figure 2.5 but randomly pick 10, 30, 50, 70 and 90 atoms from the remaining 99 materials in MERL database. Shown are aggregate statistics computed over 5 trials. We observe that the angular errors of the gradient descent approach are less than 3° for most materials even for a small, randomly-selected dictionary.

Next, we evaluate the performance of our technique with specialized dictionaries that are comprised of BRDFs from similar materials. We construct three kinds dictionaries: (i) one each for paints, fabrics, plastics, phenolics, and metals; (ii) one dictionary whose atoms are randomly selected; and (iii) a leave-one-out dictionary made of all BRDFs except the one being tested. For evaluation, we isolate 10 materials — two each for the five categories in

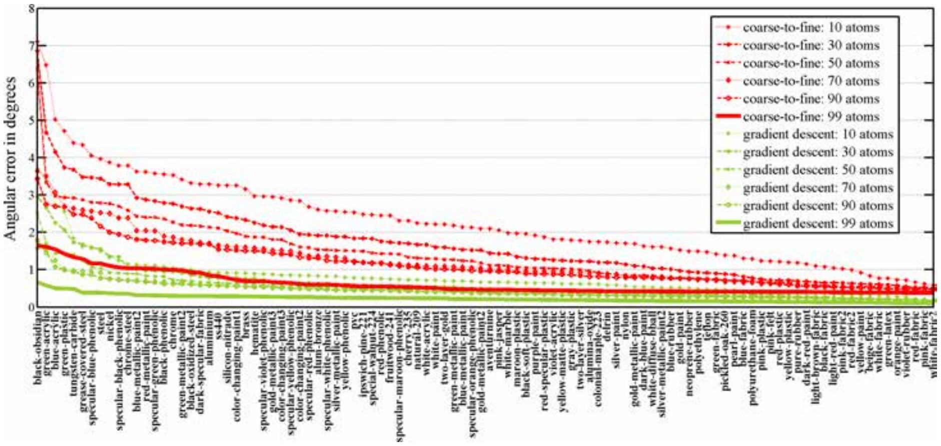


Figure 2.6: We fix the number of input images/lighting directions to 253. For different dictionary size, we compute average error over 1, 000 randomly-generated surface normals for both coarse-to-fine search (in red) and the gradient descent (in green) methods.

type (i) above — with no intersection between the training and test materials. Adopt the same setup from Figure 2.5 in terms of lighting directions and the number of input images, we evaluate the performance of the proposed technique on these 7 dictionaries in Figure 2.7. As is to be expected, a mismatch between the dictionary type and the test material produces unstable estimates. This trend is most distinct for metallic objects which have high-frequency components in their BRDFs. We also observe that the dictionaries with a mixture of materials returns the most stable performance. Finally, as expected, the leave-one-out dictionary with 99 atoms outperforms other dictionaries. This demonstrates the advantage of the proposed technique by using reference materials from a wide range.

Performance of BRDF estimation

We characterize the performance of BRDF estimation by testing on the objects with spatial invariant as well as the spatially varying BRDFs.

Given a test BRDF, we generated 100 surface normals with random orientations and rendered their appearance for 253 lighting directions. Assuming the knowledge of the true surface normals, we estimate the BRDF using the optimization in Section **??**, comparing the estimates produced by the per-pixel as well as low-rank constrained methods. We use the relative BRDF error [107] to quantify the accuracy of the estimate. Given true BRDF ρ and estimated value $\widehat{\rho}$, the relative BRDF error is given as

$$\sqrt{\sum_i w_i((\widehat{\rho}(i) - \rho(i)) \cdot \max(0, \cos(\theta_i)))^2 / \sum_i w_i}, \tag{2.8}$$

with w_i set equal to 1 for convenience.

Spatially-invariant BRDF. Figure 2.4 characterizes the average relative BRDF error for varying number of lighting directions, which is computed by averaging all 100 material BRDFs in MERL database based on the $20,000$ random generated normals. Additional results on objects with spatially-invariant BRDF can be found in the supplemental material.

Spatially-varying BRDF. Though our model qualitatively and quantitatively performs well on the homogeneous objects, objects with spatially-varying BRDF present a more challenging scenario. To illustrate this, we

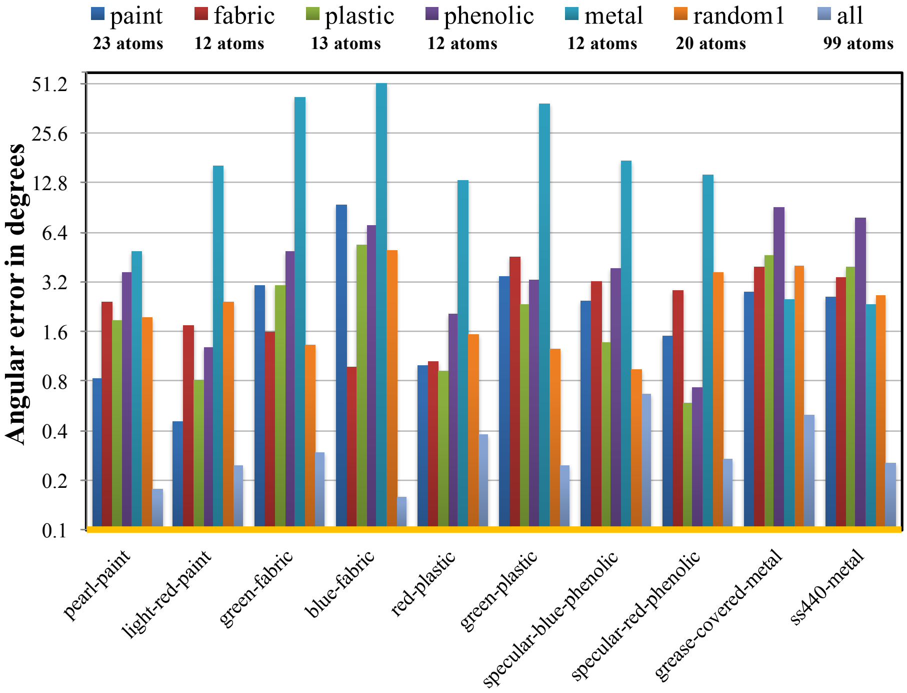


Figure 2.7: We fix the number of input images/lighting directions to 253. The numbers in the legend indicates the size of the corresponding dictionary. We observe that a mismatch in material type always leads to poor normal estimates. Shown are average errors over $1,000$ randomly-generated surface normals.

simulate an object whose SV-BRDF is constantly varying. An example is shown in Figure 2.8. We select three materials from the MERL database [103] and vary their relative abundances smoothly as shown in Figure 2.8. Now, the BRDF at each pixel in the rendered objects can be represented as a linear combination of the selected materials. In Figure 2.8, we showcase the performance of the low-rank BRDF estimation technique by rendering results obtained at different rank of the solution. We obtain solutions with varying ranks by tuning the value of β; for each value of the rank K, we show the cor-

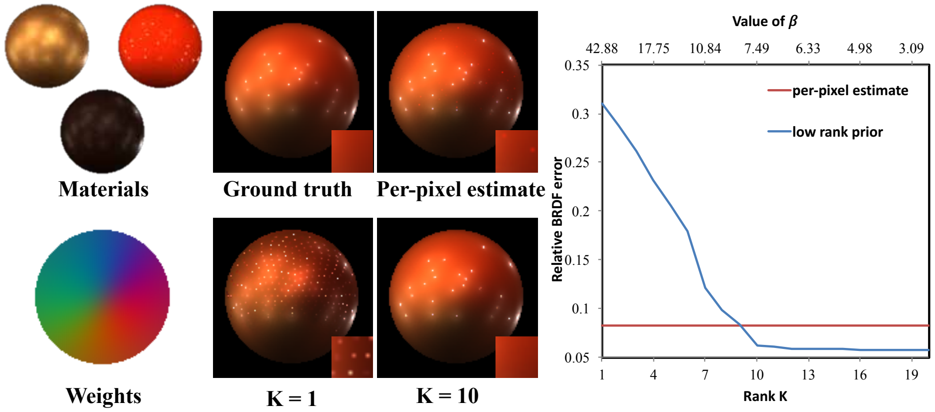


Figure 2.8: We show a spherical object whose per-pixel BRDF is a linear combination of the three materials shown. The color coded sphere shows the relative abundances of the three materials in each color channel. We show rendered images using the ground truth, the per-pixel estimate as well as the low-rank estimate for different values of rank, K. For each value of the solution rank K, we include the corresponding value of β used in the optimization at the top of the plot. Finally, we present the relative BRDF error as a function of the rank.

responding value of β at the top of the plot in Figure 2.8. The performance demonstrated in terms of both the qualitative results and quantitative measurements suggests robustness by incorporating low-rank prior. Note that the value of K used for the optimization may not be consistent with the number of the underlying materials. That is, the BRDF at each pixel, which is the linear combination of selected BRDFs, may not be uniquely described by the BRDF dictionary due to linear correlation between the atoms. This naturally introduces a larger value of K for the convergence of the relative BRDF errors.

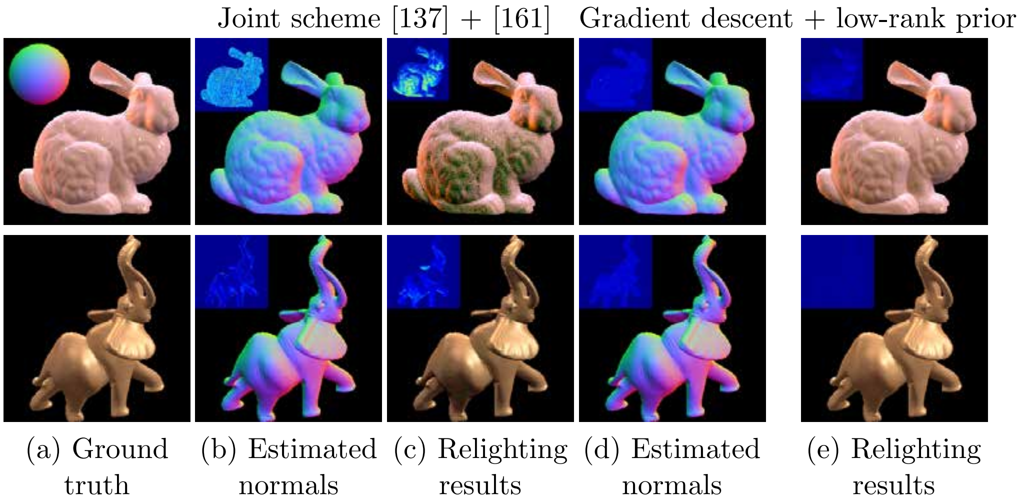


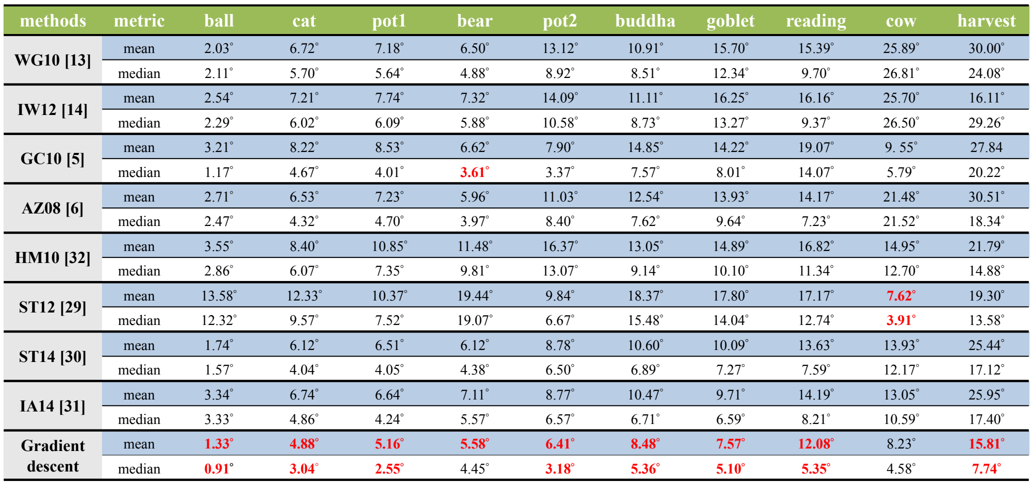
Figure 2.9: We compare on synthetic synthetic data for the proposed technique with the joint scheme of surface normals estimating using [137] and BRDF estimation using [161]. Insets on the top-left are the angular errors and euclidean intensity differences for the relighting, shown for both the joint scheme and the proposed approach. The proposed technique outperforms the competing methods in both normal estimation and novel image synthesis.

End-to-end performance characterization

To evaluate the end-to-end performance of both surface normal and BRDF estimation, we first characterize BRDF recovery using the estimated normals under varying number of light directions. Figure 2.4 characterizes the average relative BRDF errors for different number of light sources for the materials in the MERL database [103]. As seen here, the relative BRDF errors when using normal estimates of the gradient descent technique are in close proximity to errors when using the ground-truth surface normals. We also compare the relighting results for the proposed technique with the joint scheme with estimated surface normals from [137] and per-pixel BRDF fitting model from [161]. Specifically, given the surface normals estimated from [137], we perform BRDF fitting scheme as shown in [161]. We generate the test materials BRDFs by mixing materials from the MERL database and use 253 input images for the estimation. Given the estimated normals and BRDFs, we rendered the objects using the Grace Cathedral environment as the scene illumination. Figure 2.9 showcases the performance for the proposed technique and the combination of [137] and [161]. An evaluation of reconstruction error is shown in the top-left in the relighting results. We observe that the proposed technique outperforms the combination scheme for surface normal [137] and BRDF fitting [161] in terms of both visually results and quantitative measurements. While the methods [137] can produce good estimates for the surface normals and [161] can effectively address the BRDF fitting given the surface normals, they still need to solve the ill-posed problem by using priors to model the BRDFs or surface normals, making the framework fragile to the noisy estimates. In contrast, we solved for a sequence of well-defined problems, allowing for robust estimates for both surface normals and material BRDFs even noisy estimates present.

2.5.2 Real data

Real images present a layer of difficulty well beyond simulations and introduce inter-reflections, sub-surface scattering, cast shadows, and imprecise light source localization. We test the performance of our shape and BRDF recovery algorithm on a wide range of datasets. Specifically, we use datasets



methods	metric	ball	cat	pot1	bear	pot2	buddha	goblet	reading	cow	harvest
WG10 [13]	mean	2.03°	6.72°	7.18°	6.50°	13.12°	10.91°	15.70°	15.39°	25.89°	30.00°
	median	2.11°	5.70°	5.64°	4.88°	8.92°	8.51°	12.34°	9.70°	26.81°	24.08°
IW12 [14]	mean	2.54°	7.21°	7.74°	7.32°	14.09°	11.11°	16.25°	16.16°	25.70°	16.11°
	median	2.29°	6.02°	6.09°	5.88°	10.58°	8.73°	13.27°	9.37°	26.50°	29.26°
GC10 [5]	mean	3.21°	8.22°	8.53°	6.62°	7.90°	14.85°	14.22°	19.07°	9. 55°	27.84
	median	1.17°	4.67°	4.01°	**3.61°**	3.37°	7.57°	8.01°	14.07°	5.79°	20.22°
AZ08 [6]	mean	2.71°	6.53°	7.23°	5.96°	11.03°	12.54°	13.93°	14.17°	21.48°	30.51°
	median	2.47°	4.32°	4.70°	3.97°	8.40°	7.62°	9.64°	7.23°	21.52°	18.34°
HM10 [32]	mean	3.55°	8.40°	10.85°	11.48°	16.37°	13.05°	14.89°	16.82°	14.95°	21.79°
	median	2.86°	6.07°	7.35°	9.81°	13.07°	9.14°	10.10°	11.34°	12.70°	14.88°
ST12 [29]	mean	13.58°	12.33°	10.37°	19.44°	9.84°	18.37°	17.80°	17.17°	**7.62°**	19.30°
	median	12.32°	9.57°	7.52°	19.07°	6.67°	15.48°	14.04°	12.74°	**3.91°**	13.58°
ST14 [30]	mean	1.74°	6.12°	6.51°	6.12°	8.78°	10.60°	10.09°	13.63°	13.93°	25.44°
	median	1.57°	4.04°	4.05°	4.38°	6.50°	6.89°	7.27°	7.59°	12.17°	17.12°
IA14 [31]	mean	3.34°	6.74°	6.64°	7.11°	8.77°	10.47°	9.71°	14.19°	13.05°	25.95°
	median	3.33°	4.86°	4.24°	5.57°	6.57°	6.71°	6.59°	8.21°	10.59°	17.40°
Gradient descent	mean	**1.33°**	**4.88°**	**5.16°**	**5.58°**	**6.41°**	**8.48°**	**7.57°**	**12.08°**	8.23°	**15.81°**
	median	**0.91°**	**3.04°**	**2.55°**	4.45°	**3.18°**	**5.36°**	**5.10°**	**5.35°**	4.58°	**7.74°**

Figure 2.10: Shown are the mean and median of the angular errors measured in degrees for both the gradient descent method and the state-of-the-art techniques. For each object, the best performing algorithm for both mean and median angular error is marked in red. The proposed technique outperforms the benchmarked techniques in a majority of scenes. The numbers for the benchmarked algorithms are reported from [138].

from three sources — the benchmark dataset of [138], the light-stage data from [47], and the gourd from [9].

Comparisons on benchmark dataset. Figure 2.10 showcases the performance of non-Lambertian photometric stereo techniques on the benchmark dataset [138]. Each object in the database was captured with the ground truth surface normals, allowing for quantitative evaluations. For each object, we tabulate the mean and median of the angular errors for the estimated surface normals. The results for the benchmarked algorithms were done using code-base provided as part of the dataset.[3] For eight out of the ten objects, the gradient descent scheme outperforms all the methods provided as part of

[3] See `https://sites.google.com/site/photometricstereodata/`

the benchmark in both mean and median angular error.

Normal estimation. Figure 2.11 showcases shape estimation of both the coarse-to-fine search and gradient descent method, respectively, on a variety of real datasets from the USC database [47]. We use Poisson reconstruction to obtain 3D surfaces from the estimated normal maps. The results were obtained from 250 input images. From the performance of the surface, the gradient descent method provides more fine scale structures as indicated in the red rectangle (the bolt on the shoulder-plate, the bolt on the helmet, the badge), as well as remove the artifacts shown in the helmet.

BRDF estimation. Next, we showcase the performance of BRDF estimation on the `knight` scene using the surface normals estimated using the gradient descent technique. The object in the scene exhibits many unique materials (the helmet, the breast-plate, the chain, the red scabbard, to name a few) as well as significant modeling deviations (inter-reflections, cast-shadows). Figure 2.12 shows rendered photographs under natural lighting based on the USC light probes [41] for both the per-pixel and low-rank prior approaches. While the per-pixel estimates show the robustness to handle objects with complex spatial variations, it produces noisy rendering results due to insufficient observations. Incorporating the low-rank prior returns a more faithful rendition of the scene, indicating the advantages gained by pooling the information across multiple pixels.

Evaluations. Figure 2.13 showcases the performance of our algorithm on two real datasets (gourd1 and helmet). The results for the helmet were obtained from 250 input images, and the results of gourd1 were obtained from 100 input images. The recovered shape and BRDF (as visualized via rendered images) seem to be in agreement with the results in [9]; however, our algorithm is significantly simpler and employs a per-pixel algorithm that be easily parallelized. The proposed estimation framework showcases its robustness to handle objects with complex spatial varying materials and render faithful renditions under both simple and complex lighting environment. We refer the reader to the supplemental videos highlighting the relighting results.

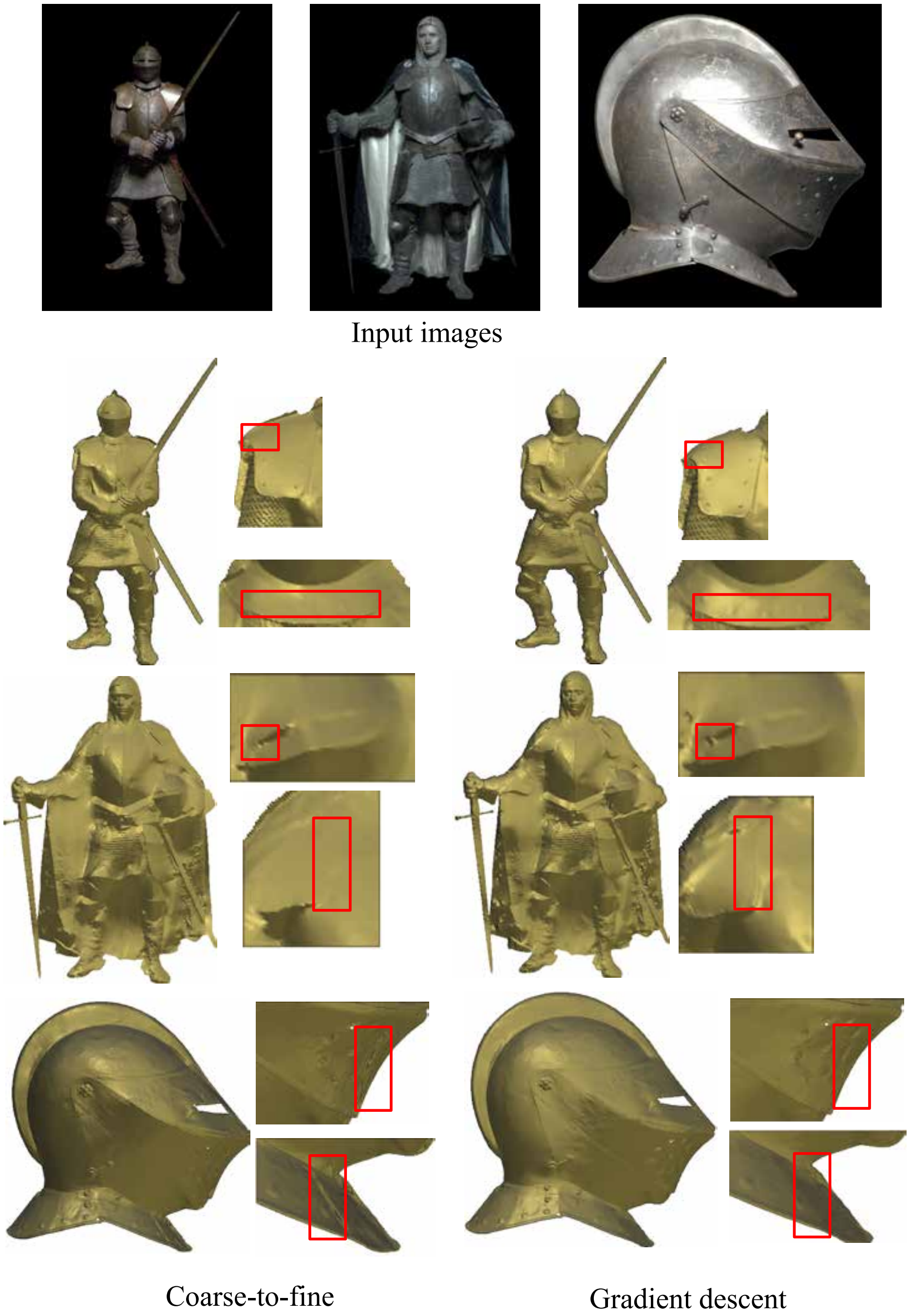


Figure 2.11: Shown are the sample images for scene from the USC light-stage database [47] and 3D surfaces obtained by using Poisson reconstruction on the estimated normal maps. We also highlight differences between the coarse-to-fine and gradient descent approaches using red boxes. We observe that the gradient descent technique is able to preserve subtle details.

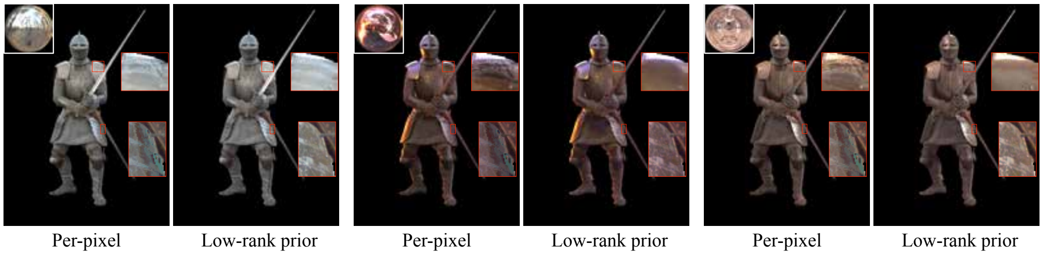


Figure 2.12: (From left to right) Shown are the rendering results under Eucalyptus Grove, Grace Cathedral and St. Peter's Basilica environment [41] for both the per-pixel and low-rank prior approaches. We also show the close-up appearance for the relighting results to highlight the improvements by incorporating the low-rank prior. Note how the shadings on the shoulder blades and thigh plates are correctly rendered by the low-rank prior.

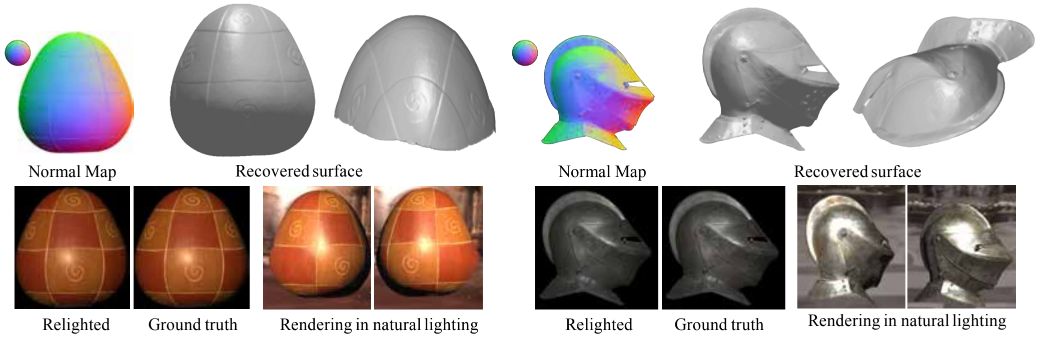


Figure 2.13: For both datasets, we show the estimated normal map in false color (top-left) and 3D surface (top-right) recovered from it. We also show the relighting results (bottom-left), ground truth under the same lighting direction (bottom-middle), and relighting under natural environment (bottom-right).

Chapter 3

Shape and reflectance estimation using a mobile device

Reflectance properties play an important role in the appearance of objects in a scene. For an opaque object, these properties are represented by the 4-D bidirectional reflectance distribution function (BRDF), which completely characterizes how a material interacts with incident light. Measuring the BRDF of a material often requires dense sampling of the 4-D space using precisely calibrated, and often prohibitively expensive, acquisition setups [40, 93, 102, 103].

More recently, researchers have looked at the problem of reflectance capture "in the wild", under relatively unconstrained conditions, and using commodity hardware. Because of the ill-posed nature of this problem, these methods rely on extra information like the presence of reference materials in the scene [125] or restrict themselves to BRDFs with stochastic, texture-like spatial variations [7].

The goal of our work is to enable the acquisition of the shape and spatially-varying BRDF (SV-BRDFs) of a wide range of real-world materials with using a practical, easy-to-deploy setup. To this end, we would like to use a mobile device — with a camera and a controllable flash — to take reflectance measurements. However, the position of the flash on these devices is fixed relative to the camera, and they are often nearly collocated. As a result, capturing images using this setup gives us only a sparse sampling of the BRDF. Even for the restricted set of isotropic materials (which are described by a 3-D BRDF), these measurements constitute only the 1-D slice of the 3-D BRDF that contains the specular lobe. We refer to this as a univariate sampling of the BRDF. The main contribution of our work is to show that such a univariate sampling of a material's appearance is, in fact, sufficient to recover per-pixel surface normals and BRDF estimates.

Real-world BRDFs can be well approximated as a linear combination of a small set of basis BRDFs [64, 103]. Based on this property, we show that while the sparse univariate samples are not sufficient by themselves, combining them with a dictionary-based prior [78] can lead to high-quality reflectance estimates. Further, we show that the parameters of many classical analytical BRDF models can be estimated purely from univariate sampling.

This is because a collocated setup samples the specular lobe of the BRDF, which plays a major role in material appearance. Thus, when constrained to take a few sparse samples of the BRDF, instead of spreading these samples across the 4-D (or a 3-D isotropic or a 2-D bivariate) space, concentrating these samples in this 1-D slice is a better way to identify the BRDF.

We use the camera and flash unit on an iPhone 6S device to scan numerous near-planar (wrt depth) targets and subsequently estimate their shape and reflectance. For each target, we capture multiple images by moving the phone. For ease of calibration of the camera/light-source orientation, we place small checkerboard patterns on the near-vicinity of the target; the acquired images are aligned via a homography estimated using these checkerboard patterns. Using the aligned images, we estimate per-pixel surface normals and SV-BRDFs using a novel, robust method based on our univariate sampling strategy. We demonstrate this robustness on a wide range of scenes with complex SV-BRDFs and further, showcase the use of the proposed BRDF acquisition technique for reflectance rendering as well as material clustering.

Contributions. Our specific contributions are as follows:

BRDF identifiability analysis. We provide a comprehensive theoretical and empirical analysis of the identifiability of BRDFs given sparse samples from a collocated setup.

Practical shape and SV-BRDF estimation. We propose a robust optimization scheme to recover per-pixel normals and BRDFs of near-planar real-world materials from images captured with a collocated setup.

Limitations. Our method is limited to near-planar samples with little depth variation. This is because we rely on a planar geometric proxy to align the multiple captured images. We assume that the images are radiometrically calibrated. The light intensity across the material sample should be uniform and significantly greater than the ambient light levels. Our method requires alignment for the input sequence. Imprecise alignment may lead to the blurry of the results. Finally, our analysis will fail for complex BRDFs like anisotropic materials and in the modeling of the Fresnel effect at grazing incidence angles.

3.1 Prior work

Direct BRDF measurement. The BRDF is a function of four angles, two each for incident and outgoing directions, and classical BRDF measurement approaches [43, 102, 103] sample this 4D space by capturing images under varying lighting and viewing directions. Densely sampled measured BRDFs can provide faithful renditions of material appearance, but require specialized acquisition setups to capture large numbers of images.

Photometric stereo methods. Photometric stereo methods aim to recover shape from images captured with a fixed camera and varying lighting.

While originally proposed for materials with known reflectance [140, 153], they have been extended to jointly infer shape and reflectance properties. This is done by using low-dimensional parametric reflectance models such as the isotropic Ward model [64], or directional statistics BRDF model [111, 112, 114]. Alternatively, the form of the reflectance function is constrained, typically by assuming that the BRDF is isotropic. Romeiro et al. [129] show that isotropic BRDFs are well-approximated by a 2-D bivariate representation and use this to recover BRDF from a single image of a known shape under known illumination. The bivariate representation has been used for shape and SVBRDF estimation from multiple images [9, 139], and blind reflectance recovery from a single image of a known shape [131]. Chandraker et al. [33–35] leverage motion cues to recover shape and reflectance for objects with dichromatic BRDFs. While more general than direct BRDF capture, these methods rely on restricted setups (calibrated, distant lighting and cameras) and/or extra information (known geometry, homogenous BRDFs). Our goal is to capture general SV-BRDFs using a light-weight capture setup.

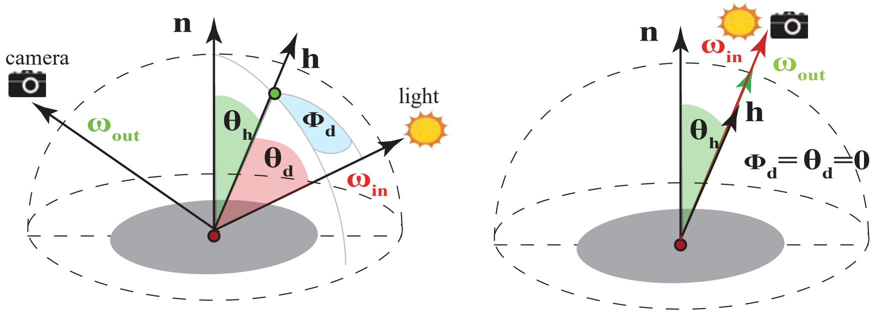
Optimal BRDF sampling. Nielsen et al. [108] address the problem of identifying the optimal set of reflectance measurements required to recover a BRDF. This idea is further extended by Xu et al. [156] to consider near-field measurements. These papers show that a small set of images – in some cases, even two – are sufficient to estimate a BRDF. However, they are restricted to homogeneous materials and the nature of these measurements requires two pre-calibrated cameras and light sources. In contrast, we seek to recover SV-BRDFs using commodity hardware, and we demonstrate that this is possible using a collocated setup by enforcing a dictionary-based prior on the reconstruction.

BRDF acquisition using commodity devices. Higo et al. [73] capture images with a hand-held camera with an attached point light source and use a combination of near-light Photometric Stereo and multi-view stereo to reconstruct roughly Lambertian objects. Ren et al. [125] show that SV-BRDFs can be acquired using a fixed camera and a moving hand-held source by placing reference material tiles in the scene. While their results are impressive, the use of reference materials makes this setup less practical in real-world situations. Aittala et al. [7] propose to estimate SVBRDFs and normal maps from flash/no-flash image pairs captured using mobile devices. They extend this work to a single image using neural network-based texture features [6]. However, these methods are restricted to stationary texture-like SVBRDFs and are aimed at reproducing plausible texture variations rather than accurate measured BRDF reconstruction. Riviere et al. [127] propose two prototypes using a mobile camera-flash or an LCD panel for reflectance capture. Their mobile camera solution can only handle rough specular surfaces and their shape and BRDF estimates are largely based on heuristics. In contrast, we can handle a wider range of materials because of our robust dictionary-based shape and reflectance estimation.

3.2 Univariate sampling of BRDFs

While arbitrary BRDFs are 4D functions of reflectance, many real-world materials are isotropic, in that, their BRDF is invariant to joint rotations of the incident and outgoing directions about the surface normal. The BRDF of such isotropic materials can be represented with a three-angle coordinates system, often using the half-angle parameterization [132] that is defined as follows. Given the surface normal $\mathbf{n}$, the incident direction $\boldsymbol{\omega}_i$ and the outgoing direction $\boldsymbol{\omega}_o$ — all unit-norm vectors — we first compute the half-angle $\mathbf{h} = (\boldsymbol{\omega}_i + \boldsymbol{\omega}_o)/2$. Next we define (θ_h, ϕ_h) as the elevation and azimuth, respectively, of the half-angle with respect to the surface normal, and (θ_d, ϕ_d) as the elevation and azimuth, respectively, of the outgoing direction with respect to the half-angle (see Figure 3.1). An isotropic BRDF, represented as $\rho(\theta_h, \theta_d, \phi_d)$, is represented as a function over θ_h, θ_d, and ϕ_d with $\theta_h, \theta_d \in [0, \pi/2)$ and $\phi_d \in [0, \pi)$. A subsequent reduction in dimensionality is provided by bivariate models [129] that further assume that the BRDF is invariant to changes in ϕ_d, and hence, the resulting reflectance is simply a function of θ_h and θ_d.

Collocated systems and univariate sampling. When the light source and the camera are collocated, then the incident and outgoing directions are the same, i.e., $\boldsymbol{\omega}_i = \boldsymbol{\omega}_o = \mathbf{h}$. Hence, $\theta_d = \phi_d = 0^\circ$. Hence, any sampling of the BRDF is purely a function of θ_h. We refer to this paradigm as univariate sampling. Further, when there is a small, but fixed, offset between the



Half-angle representation

Collocated setup

Figure 3.1: Schematic of half-angle BRDF representation with respect to $(\theta_h, \theta_d, \phi_d)$ and univariate sampling on θ_h.

light source and camera, then θ_d and ϕ_d are no longer zero, but are known constants independent of θ_h and ϕ_h, and hence can be pre-calibrated.

An important question to resolve upfront is whether univariate sampling can provide sufficiently rich measurements to be able to capture salient features of the measured BRDF, as well as enable stable reconstructions of the BRDF. We address this in two different ways. First, in Section 3.1, we show that the parameters of many analytical BRDF models are identifiable from noiseless univariate samples. Second, in Section 3.2, we provide empirical results characterizing accuracy of BRDFs, estimated from univariate samples.

3.2.1 Identifiability under univariate sampling

We now address the question of identifiability of BRDFs from univariate samples, i.e., in the absence of noise, can there exist two distinct BRDFs that produce the same set of univariate samples? The answer is a resounding yes,

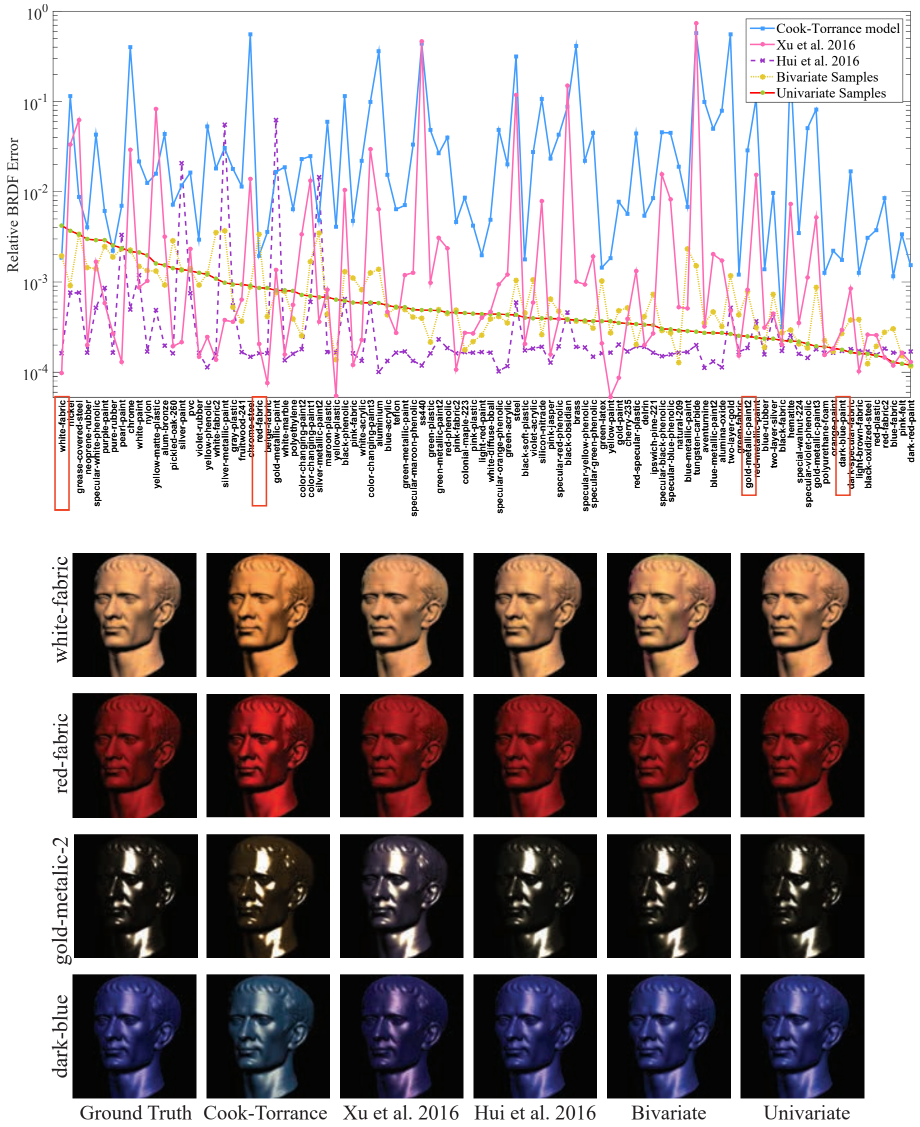


Figure 3.2: We demonstrate the performance of univariate sampling against a number of other sampling strategies. Shown are the relative BRDF errors of the reconstructed BRDF on MERL database for our technique, the data-driven method of Hui et al. [78], the bivariate model [129], the parametric model of Cook-Torrance, and optimal sampling model of Xu et al. [156]. We also compare against a 2D sampling strategy that we refer to as "bivariate sampling" that provides samples in a bivariate BRDF space (θ_h, θ_d). We observe that the method of Hui et al. [78] returns the best performance on an average. However, the univariate sampling with the proposed prior is able to compete against most of the state-of-the-art methods, both quantitatively as well as qualitatively.

if we do not further constrain the BRDF in some meaningful way. We do so by restricting ourselves to popular parametric BRDF models, and show that the parameters of the models are identifiable from noiseless univariate samples. Given the space constraints, we show this for the Cook-Torrance model [39] and provide justifications for other models including the Blinn-Phong, isotropic Ward and the Ashikhmin-Shirley model.

Proposition. The parameters of the Cook-Torrance model are identifiable from noiseless univariate measurements.

Proof. BRDF measurements under the Cook-Torrance model are dependent on two parameters: m and F_0. Under univariate sampling, the BRDF can be written as:

$$\rho(\theta_h) = \rho_d + \frac{(1-\rho_d)DGF}{\pi\ (\mathbf{n}^\top \mathbf{l})\ (\mathbf{n}^\top \mathbf{v})} = \rho_d + \frac{(1-\rho_d)DGF_0}{\pi \cos^2\theta_h} \tag{3.1}$$

where

$$D = \frac{e^{-\tan^2\theta_h/m^2}}{m^2\cos^4\theta_h}, \quad G = \min(1, 2\cos^2\theta_h).$$

The term G is purely a function of θ_h and does not depend on any parameters, i.e. F_0 and m. Note that the Fresnel term, F, reduces to a constant F_0 for a collocated setup.

First, we observe that $\rho_d = \rho(\pi/2)$.[1] Second, we can now rearrange (3.1)

[1] In practice, due to fore-shortening, we cannot make an observation at $\theta_h = \pi/2$; however, this can easily be handled by sampling the BRDF at values close to $\pi/2$ and predicting the limiting value.

to the following expression:

$$\log \frac{\pi(\rho(\theta_h) - \rho_d) \cos^6 \theta_h}{(1 - \rho_d)G} = \log \left(\frac{F_0}{m^2} \right) - \frac{\tan^2 \theta_h}{m^2} \tag{3.2}$$

Note that we have complete knowledge of the LHS term in (3.2). Further, if we plot the LHS as a function of $\tan^2 \theta_h$, then the resulting plot is expected to be a straight line whose slope is $-1/m^2$ and whose intercept is $\log(F_0/m^2)$. Hence, we can identify all parameters of the model from the univariate measurements. □

Blinn-Phong model. BRDF measurements under the Blinn-Phong model and univariate sampling (i.e., $\theta_d = \phi_d = 0$) can be written as:

$$\rho(\theta_h) = \rho_d + \rho_s \frac{\beta + 2}{2\pi} \cos^\beta \theta_h,$$

where ρ_d, ρ_s and β are the parameters defining the model. Given ρ_d, we can write

$$\log(\rho(\theta_h) - \rho_d) = \log \rho_s + \log(\beta + 2) - \log(2\pi) + \beta \log \cos \theta_h.$$

Hence, if we plot $\log(\rho(\theta_h) - \rho_d)$ as a function $\log \cos \theta_h$, then the resulting plot is a straight line whose slope is β and intercept is $\log \rho_s + \log(\beta + 2) - \log(2\pi)$. Hence, we can recover all three parameters of the model.

Isotropic Ward model. BRDF measurements under the Isotropic Ward model and univariate sampling (i.e., $\theta_d = \phi_d = 0$) can be written as:

$$\rho(\theta_h) = \rho_d + \frac{\rho_s}{\cos\theta_h} \frac{\exp(-\tan^2\theta_h/\beta^2)}{4\pi\beta^2},$$

where ρ_d, ρ_s and β are the parameters defining the model. First, we observe that $\rho_d = \rho(\pi/2)$. Second, given ρ_d, we can write

$$\log(\rho(\theta_h) - \rho_d) = \log\left(\frac{\rho_s}{4\pi\beta^2}\right) - \log(\cos\theta_h) - \frac{\tan^2\theta_h}{\beta^2}.$$

Equivalently,

$$\log(\rho(\theta_h) - \rho_d) + \log(\cos\theta_h) = \log\left(\frac{\rho_s}{4\pi\beta^2}\right) - \frac{\tan^2\theta_h}{\beta^2}.$$

If we plot the LHS expression as a function of $\tan^2(\theta_h)$, then we expect a straight line whose slope is $1/\beta^2$ and intercept is $\log\rho_s - \log(4\pi\beta^2)$, from which we can identify both β and ρ_s. More specifically, we can identify these parameters from values of $\rho(\theta_h)$ at two distinct values of θ_h. The proof for the Ashikhmin-Shirley model follows very closely the one we described above for the Cook-Torrance model.

3.2.2 Empirical validation

Next, we show that BRDFs can be estimated reliably from univariate measurements. Univariate samples provide a highly under-determined set of measurements and hence, recovering BRDFs from them requires the use of strong reflectance priors. We use a dictionary-based model for this purpose, borrowing an idea proposed recently in Hui et al. [78, 79].

Dictionary-based BRDF models. There have been many approaches [9, 64, 156] that model the BRDF at each pixel to lie in the non-negative span of a set of exemplar BRDFs, that we refer to as a dictionary. A dictionary D is simply a collection of exemplar BRDFs, often grouped together as a matrix $D = [\rho_1, \rho_2, \ldots, \rho_M]$, where each column is the BRDF of a measured material. Given D, we represent a BRDF ρ as:

$$\rho = D\mathbf{c}, \quad \mathbf{c} \geq 0.$$

Instead of estimating the high-dimensional vector ρ, we only need to estimate the abundances $\mathbf{c}$, whose dimension is proportional to the number of materials in the dictionary. Following Hui et al. [78], we further assume that $\mathbf{c}$ is sparse, suggesting that BRDF is the linear combination of a few dictionary atoms.

BRDF recovery. Univariate sampling measurements can be written as follows:

$$\begin{aligned} y(\theta_h) &= S(\theta_h)\rho + \eta \\ &= S(\theta_h)D\mathbf{c} + \eta, \end{aligned}$$

where $S(\theta_h)$ is the linear sampling operator that extracts the value at the input BRDF at $(\theta_h, 0, 0)$ and η is the measurement noise. Given M samples, corresponding to half-angle elevations in the set $\{\theta_h^1, \ldots, \theta_h^M\}$, we can compute the coefficients $\mathbf{c}$ by solving for the problem as

$$\widehat{\mathbf{c}} = \arg\min_{\mathbf{c} \geq 0} \sum_{i=1}^{M} \|y(\theta_h^i) - S(\theta_h^i) D\mathbf{c}\|_2^2 + \lambda \|\mathbf{c}\|_1. \tag{3.3}$$

We can now obtain the BRDF estimate $\widehat{\rho} = D\widehat{\mathbf{c}}$. The procedure illustrated above is from the dictionary-based modeling of BRDFs in Hui et al. [78], adapted to the univariate sampling scenario.

Comparisons. We evaluate the performance of the reconstruction technique with the state-of-the-art methods on the entire MERL database using a leave-one-out scheme. In particular, we compare against the parametric model of Cook-Torrance, the optimal sampling method in Xu et al. [156], and the isotropic sampling in Hui et al. [78]. For the Cook-Torrance model, we used the parameters reported in [107] — these parameters were optimized over the entire BRDF. For Hui et al. [78], we fix the surface normal at the north pole $[0,\ 0,\ 1]^\top$ and randomly sample the isotropic BRDF space for 20 combinations of lighting/view directions and reconstruct using a dictionary-based prior. For Xu et al. [156], we used the 20 optimal BRDF entries indicated in their work. For the univariate sampling, we randomly sample the θ_h axis of the BRDFs and collect 20 samples with collocated lighting and view direction. Similarly, we also sample the bivariate BRDF space spanned by θ_h

and θ_d with the 20 lighting/view combinations; we refer to this as bivariate sampling and use the same recovery algorithm as with univariate samples. For the results of Hui et al [78], univariate, and bivariate sampling, we perform 5 different random trials and report the average errors in Figure 3.2. The relative BRDF errors for these methods are shown in Figure 3.2, where we observe that univariate sampling is quite competitive to state-of-the-art models.

Varying number of input images. Figure 3.3 characterizes the errors in surface normal as well as the BRDF estimation for varying number of input images, or equivalently, the number of samples for the half angles. For the normal estimation, we report the average angular error for the univariate sampling. The average angular error is computed by randomly generating 100 normals per material and varying across all 100 material BRDFs in the database. To characterize the performance for BRDF estimation, we assume the knowledge of true surface normal and report the average relative BRDF error for varying number of images. As noted from Figure 3.3, both the angular errors and relative BRDF errors degrade gracefully with a smaller number of images.

BRDF estimation against ground truth. In this experiment, we evaluate how accurately we can reconstruct a BRDF given only univariate samples using a dictionary-based prior. We assume a collocated setup, i.e. $\theta_d = 0°$, set $\phi_h, \phi_d = 0°$, and sample a chosen MERL BRDF at different values

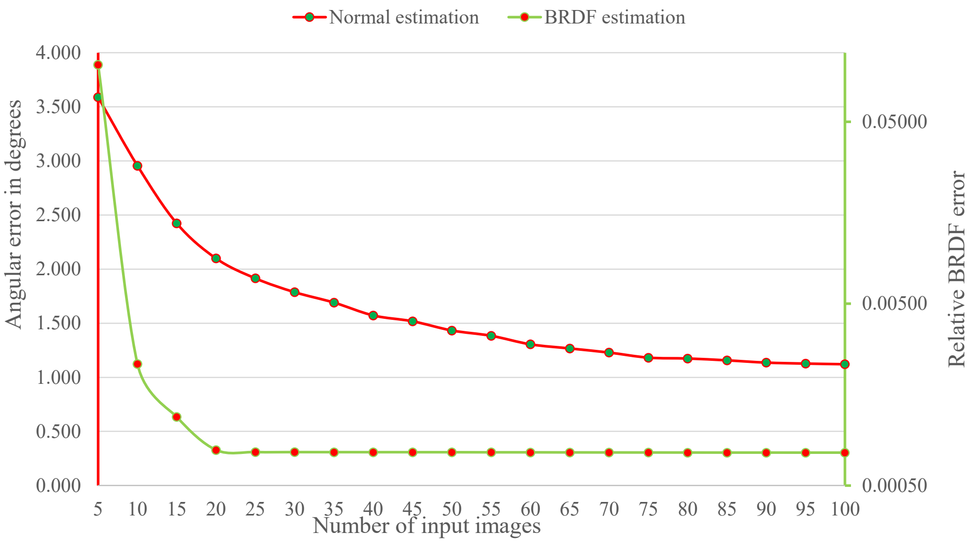


Figure 3.3: We estimate the angular errors as well as the relative BRDF errors for our proposed method on the MERL database. The plots were obtained by averaging across all 100 BRDFs in the MERL database and 100 randomly-generated normals per material.

of θ_h to obtain the univariate measurements. In Figure 3.4, we visualize the original BRDF (parameterized by (θ_d, θ_h)) as well as 3 different 1-D slices corresponding to $\theta_d = (0°, 15°, 30°)$. We then reconstruct the BRDF using a dictionary composed of the remaining 99 BRDFs in the MERL database, using only the samples corresponding to $\theta_d = 0°$. Figure 3.4 also visualizes the reconstructed BRDFs, as well as 1-D slices of these reconstructions at $\theta_d = (0°, 15°, 30°)$. As can be seen here, in spite of estimating the BRDF from only one slice of samples, the reconstructed results closely match the ground truth BRDFs even at other angles as indicated by the measured RMSE values.

Varying baseline between camera and flash. Smartphones and tablets have different relative positioning of their cameras and flash units. Hence, it is important to characterize the stability of our reflectance estimation technique for varying baseline. When the object is at an approximately fixed distance from the device, changes in baseline can be modeled as changes in θ_d. Figure 3.5 showcases this by characterizing BRDF estimation errors as a function of θ_d. For each θ_d, we compute average error over 100 materials in MERL database. Note that the performance remains stable for θ_d less than 65 degree; this is sufficient to capture the operating scenario underlying a wide range of mobile devices. Beyond 65 degrees, the univariate sampling completely misses the specular lobe which results in poor performance in estimating the BRDF.

In addition, we ability of our method to handle a wide range of materials, by rendering spheres lit by point lights for different values of θ_d (see Figures 3.6). Renderings with our reconstructed BRDFs closely match those produced using ground truth BRDF data, demonstrating that our method is able to produce realistic results for a majority of isotropic materials in the MERL database.

Remark. For many materials, the univariate sampling outperforms competing methods that sample in the bivariate space (θ_h and θ_d) or the isotropic space (θ_h, θ_d and ϕ_d). Given that we enforced a measurement budget for all methods, univariate sampling enjoys a denser sampling of the specular

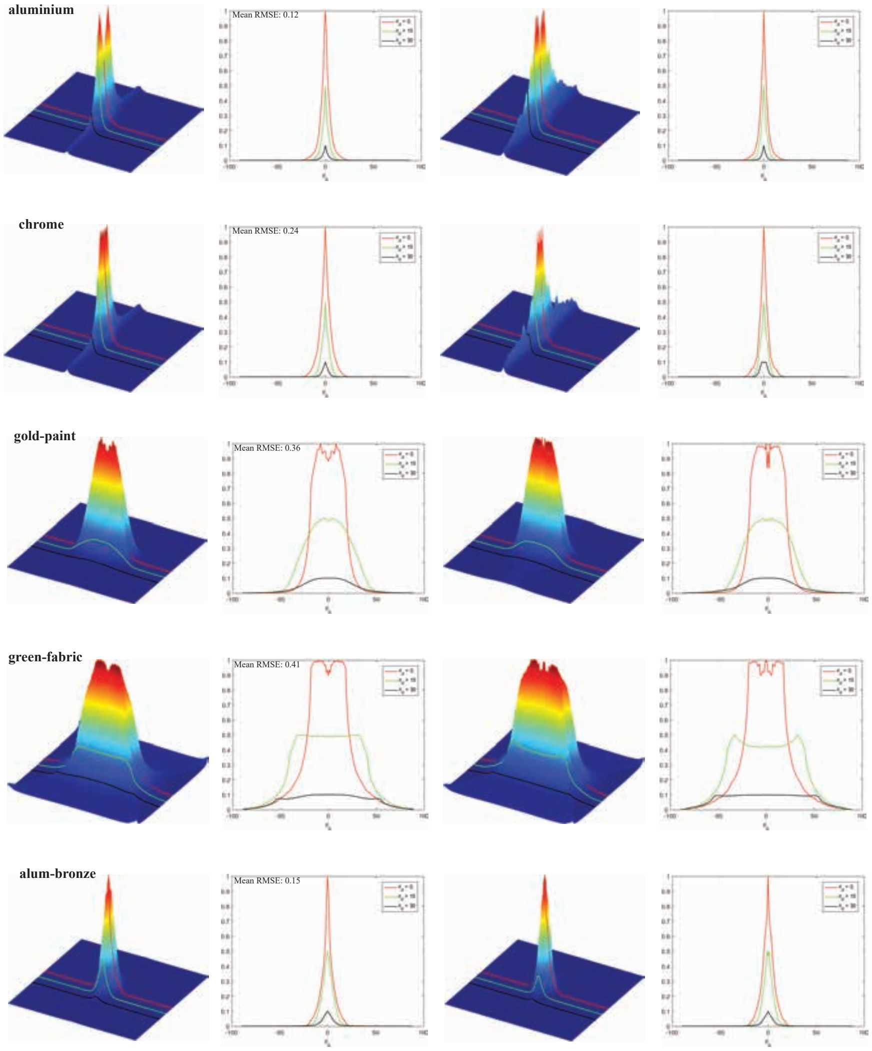


Figure 3.4: On the left, we visualize ground truth MERL BRDFs on the 2-D plane parameterized by (θ_d, θ_h) (for $\phi_d = 0°$), as well as 3 1-D slices corresponding to $\theta_d = (0°, 15°, 30°)$. We reconstruct the BRDF using only the collocated univariate samples, i.e., $\theta_d = 0°$, and visualize it on the right. As can be seen here, by using the univariate measurements only, we can reconstruct the BRDF at other θ_d quite accurately, as indicated by the mean RMSE shown in the top-left of each plot. We normalize the BRDF ρ by $\widehat{\rho} = \frac{\rho - \min(\rho)}{\max(\rho) - \min(\rho)}$ and plot the curve with different θ_d.

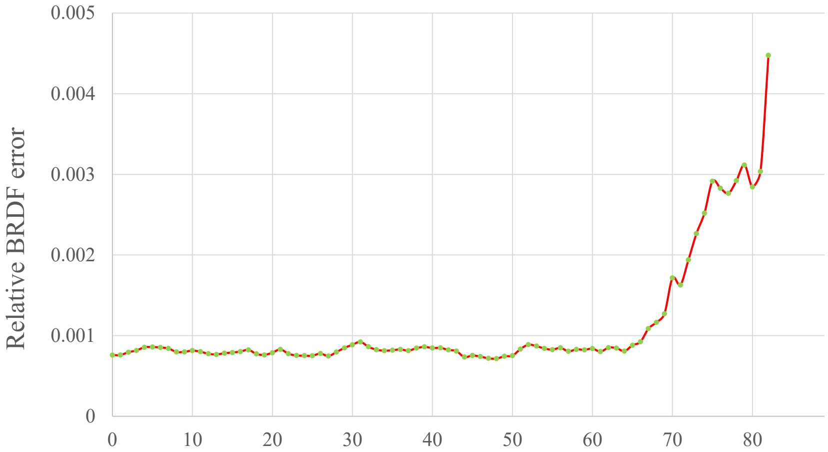


Figure 3.5: We plot the approximation accuracy in terms of relative BRDF errors by varying θ_d in degrees. For each θ_d, we compute average error over 100 materials in MERL database.

lobe. However, as we increase the number of measurements, univariate sampling has diminishing returns in reconstruction performance while competing methods that perform full sampling as well as bivariate sampling continue to observe significant gains. Our empirical evaluation also indicates that BRDFs of real-world materials are highly redundant and that the univariate sampling of an isotropic BRDF for $\theta_d = 0$ is often sufficient for high-quality reconstructions. This hypothesis is similar in spirit to bi-polynomial BRDF model introduced by Shi et al. [137], providing the BRDF as the product of two univariate function over θ_d and θ_h, respectively.

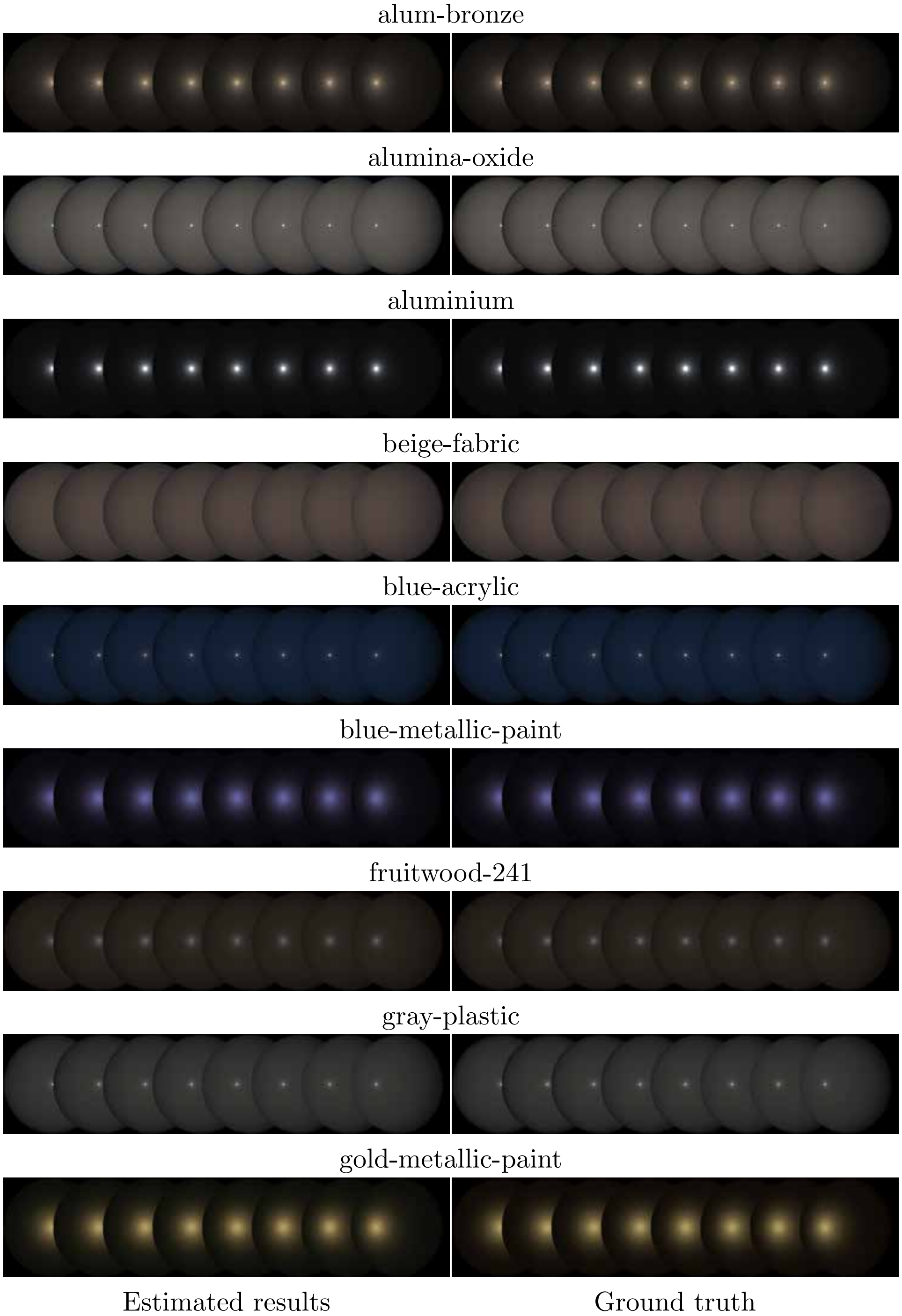


Figure 3.6: We visualize the ground truth MERL BRDF data and our univariate sample-based reconstruction by rendering these BRDFs on spheres lit by point light sources for different values of θ_d. Our rendered materials are visually almost identical to the ground truth BRDFs, indicating the accuracy of our reconstruction.

3.3 Shape and reflectance estimation under univariate sampling

Acquisition setup and calibration. Our imaging setup consists of a nearly-collocated camera and light source, we assume the intrinsic matrix of the camera is known via a one-time pre-calibration. We acquire Q (typically, about 100) images at different viewpoints of a target. We assume the target is nearly planar, mainly for ease of registering the images across different viewpoints using homography-based methods. For each view, we use the four checker board patterns attached to the corners of the target to compute the homography. The checker board patterns also allow us to compensate the lighting variations within each captured image. Using the homography, we align pixels across different images and find world coordinates of all pixels. We now have a stack of intensity observations under known lighting and viewing directions for each pixel.

Problem statement. Given the aligned images, we can formulate the objective function that incorporates both surface normal and BRDF at pixel $\mathbf{p}$ as

$$\{\widehat{\mathbf{n}}_{\mathbf{p}}, \widehat{\mathbf{c}}_{\mathbf{p}}\} = \underset{\mathbf{c}\geq 0,\mathbf{n}}{\arg\min} \|\mathbf{I}_{\mathbf{p}} - B(\mathbf{n}, \mathbf{l}_{\mathbf{p}}, \mathbf{v}_{\mathbf{p}})\mathbf{c}\|_2^2 + \lambda\|\mathbf{c}\|_1 \tag{3.4}$$

where $\mathbf{I}_{\mathbf{p}} \in \mathbb{R}^Q$ denotes the image intensities observed at pixel $\mathbf{p}$ after alignment, $\mathbf{l}_{\mathbf{p}}$ and $\mathbf{v}_{\mathbf{p}}$ are the lighting and viewing directions for Q collected images, i.e. $\mathbf{l_p} = [l_{\mathbf{p}}^1, l_{\mathbf{p}}^2, \dots l_{\mathbf{p}}^Q]$, $\mathbf{v_p} = [v_{\mathbf{p}}^1, v_{\mathbf{p}}^2, \dots v_{\mathbf{p}}^Q]$. Note that $\mathbf{l_p}$ and $\mathbf{v_p}$ are known via the calibration. The term $B(\mathbf{n}, \mathbf{l_p}, \mathbf{v_p})$, an $Q \times M$ matrix, is given as

$$B(\mathbf{n}, \mathbf{l_p}, \mathbf{v_p}) = S(\mathbf{n}, \mathbf{l_p}, \mathbf{v_p})D,$$

where S has Q rows and a number of columns equal to the dimensionality of the BRDFs; here, S encodes the shading term as well as sampling of the BRDF. The estimates of the surface normal $\mathbf{n_p}$ and the abundance $\mathbf{c_p}$ amount to solving a quadratic cost function with ℓ_1-norm constraint.

Identifying BRDF exemplars. For computational efficiency, we enforce the sparsity prior on the abundances by first identifying a compact set of BRDF exemplars for a material sample. Specifically, we solve for the abundances at each pixel via (3.4) with initialized flat surface and sum the abundances across all pixels. Now, we obtain the summed result $\boldsymbol{\mathcal{C}} \in \mathbb{R}^M$, where M is the number of atoms in the dictionary. We empirically observe that only few atoms in $\boldsymbol{\mathcal{C}}$ have large values while the remaining entries are close to zero, which is consistent with the observation in [9, 78].

We retain only the K (in our case $K = 10$) BRDFs with the highest values of $\boldsymbol{\mathcal{C}}$ as our compact set of BRDF exemplars. This obviates the need for the sparsity constraint in subsequent iterations, thus speeding up computation. We denote $\widehat{B}$ as the dictionary with columns that corresponds to the exemplar set of atoms. We now solve for the normals and the coefficients:

$$\{\widehat{\mathbf{n_p}}, \widehat{\mathbf{c}}_\mathbf{p}\} = \underset{\mathbf{c}\geq 0, \mathbf{n}}{\arg\min} \|\mathbf{I_p} - \widehat{B}(\mathbf{n}, \mathbf{l_p}, \mathbf{v_p})\mathbf{c}\|_2^2. \tag{3.5}$$

Surface normal and SV-BRDF estimation. Given the initial estimate of $\mathbf{c}^{(0)}$ from flat surface and $\widehat{B}$, we use an iterative local search to solve for the surface normals. Specifically, we build a 2D grid with respect to the elevation and azimuth angles, and search in the grid for the normals which can best describe the intensity profile. In the first iteration, we initialize all the surface normals pointing toward the north pole, i.e., a flat surface, and solve for the abundances $\widehat{\mathbf{c}}^0$ via (3.4). Now, at the j-th iteration, we have normal estimate $\widehat{\mathbf{n}}_\mathbf{p}^{(j-1)}$ with elevation angle $\theta_\mathbf{p}^{(j-1)}$ and azimuth angle $\phi_\mathbf{p}^{(j-1)}$. The 2D grid for the (j)-th iteration is constructed as

$$\mathcal{N}^{(j)} = \{(\tilde{\theta}, \tilde{\phi}) | |\tilde{\theta} - \theta_\mathbf{p}^{(j-1)}| \leq \mathcal{T}_\theta, |\tilde{\phi} - \phi_\mathbf{p}^{(j-1)}| \leq \mathcal{T}_\phi\},$$

where $\mathcal{T}_\theta$ and $\mathcal{T}_\phi$ are the thresholds to determine the cardinality of the candidate set. We can incorporate a coarse-to-fine search by specifying different values for $\mathcal{T}_\theta$ and $\mathcal{T}_\phi$, where $\mathcal{T}_\theta$ is varying from 5 to 0.1 degree while $\mathcal{T}_\phi$ is changing from 50 to 1 degrees. For each element in $\mathcal{N}^{(j)}$, the candidate surface normal is computed as

$$\tilde{\mathbf{n}} = [\sin(\tilde{\theta})\cos(\tilde{\phi}),\ \sin(\tilde{\theta})\sin(\tilde{\phi}),\ \cos(\tilde{\theta})]$$

The estimate of the surface normal at a pixel $\mathbf{p}$ is given as

$$\widehat{\mathbf{n}}_{\mathbf{p}}^{(j)} = \arg \min_{\mathbf{n}_{\mathbf{p}} \in \mathcal{N}^{(j)}} \|\mathbf{I}_{\mathbf{p}} - \widehat{B}(\mathbf{n}_{\mathbf{p}}, \mathbf{l}_{\mathbf{p}}, \mathbf{v}_{\mathbf{p}}) \mathbf{c}_{\mathbf{p}}^{(j-1)}\|_2^2. \tag{3.6}$$

This is solved by scanning over all the elements in $\mathcal{N}^j$. Note that $\mathbf{c}_{\mathbf{p}}$ has kept fixed with the values from the $(j-1)$-th iteration. Once we obtain $\widehat{\mathbf{n}}_{\mathbf{p}}^{(j)}$, we update the coefficients $\mathbf{c}_{\mathbf{p}}$ by solving

$$\widehat{\mathbf{c}}_{\mathbf{p}}^{(j)} = \arg \min_{\mathbf{c}_{\mathbf{p}}} \|\mathbf{I}_{\mathbf{p}} - \widehat{B}(\widehat{\mathbf{n}}_{\mathbf{p}}^{(j)}, \mathbf{l}_{\mathbf{p}}, \mathbf{v}_{\mathbf{p}}) \mathbf{c}_{\mathbf{p}}\|_2^2 \quad s.t.\ \mathbf{c}_{\mathbf{p}} \geq 0. \tag{3.7}$$

The algorithm typically converges within 10 iterations. The ultimate estimate of BRDF at each pixel is $\widehat{\rho}_{\mathbf{p}} = \widehat{D}\widehat{\mathbf{c}}_{\mathbf{p}}^{(J)}$, where $\widehat{D}$ corresponds to the selected columns for $\widehat{B}$ and J denotes for the number of iterations.

3.4 Results and Applications

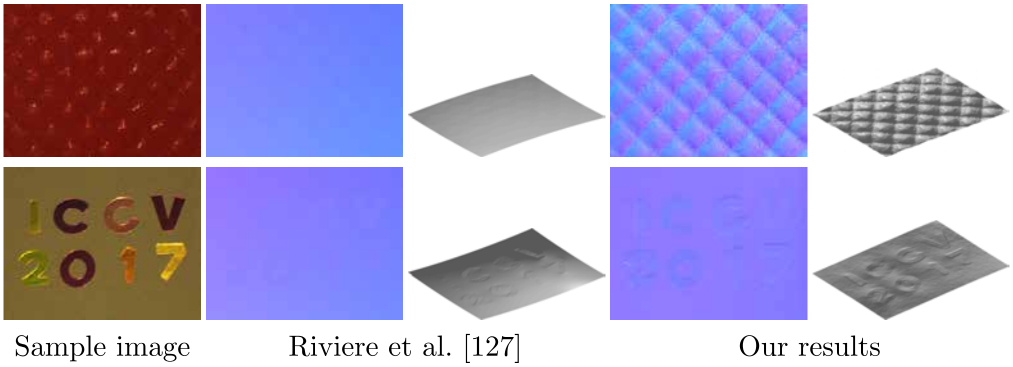
In this section, we characterize the performance of our technique on a wide range of real-world scenes captured with iPhone 6s for a variety of tasks. We fix the target sample and move the phone while capturing images under the phone's flash illumination. The images were captured with 2016 × 1512 pixels and we crop the regions with the target object for shape and BRDF estimation. We recover the per-pixel BRDFs with 1 degree for each angle in BRDF space, which leads to a $90 \times 90 \times 180 = 1,458,000$ dimensional vector. We direct the reader to the accompanying supplementary material for more results, comparisons, and analysis.

3.4.1 Shape and Reflectance Estimation

We process the captured images using the technique detailed in Section **??** to recover per-pixel surface normals and SV-BRDFs. We integrate the estimated normals using Poisson reconstruction [5] to obtain the 3D surface.

Shape estimation. To evaluate the performance of our shape estimation, we compare against the work of Riviere et al. [127], who use a similar mobile camera-based setup. While we model near-field camera and lighting, they assume that the camera and light are distant. In addition, their reflectance estimation is based on image heuristics, unlike our optimization-based framework with a BRDF prior. As demonstrated in Figure 3.7, our technique recovers more fine scale structures than [127]. In addition, our technique successfully separates reflectance effects from geometry, and as a result our reconstructions are largely planar. In contrast, their BRDF errors leak into the shape estimates leading to deviations from the planar structure of the samples. More comparisons with [127] on both real and synthetic scenes can be found in supplementary material.

Reflectance capture. Figure 3.8 illustrates the performance of our method on datasets captured using an iPhone 6s. These four datasets — `leaf`, `leather`, `fur` and `characters` — have 123, 126, 70, and 138 input im-



Sample image Riviere et al. [127] Our results

Figure 3.7: We compare our performance on surface normal estimation with Riviere et al. [127] on two datasets. Shown are (left-right) one sample image, estimated normals and recovered 3D shape via Poisson reconstruction. Please note that our reconstructions, like the actual samples, are close to planar and contain more fine-scale detail.

ages, respectively. For each dataset, we show the estimated surface normals and recovered 3D shape under different viewpoints. The surface reconstructions show that we can recover fine-scale geometric details like yarn threads and leather patterns, even for samples with complex BRDFs. While we use a large number of input images to produce the results, our experience is that the performance degrades gracefully with a smaller number of images. We direct the reader to the supplementary materials where we include the BRDF/normal estimation error as a function of number of images on the synthetic dataset.

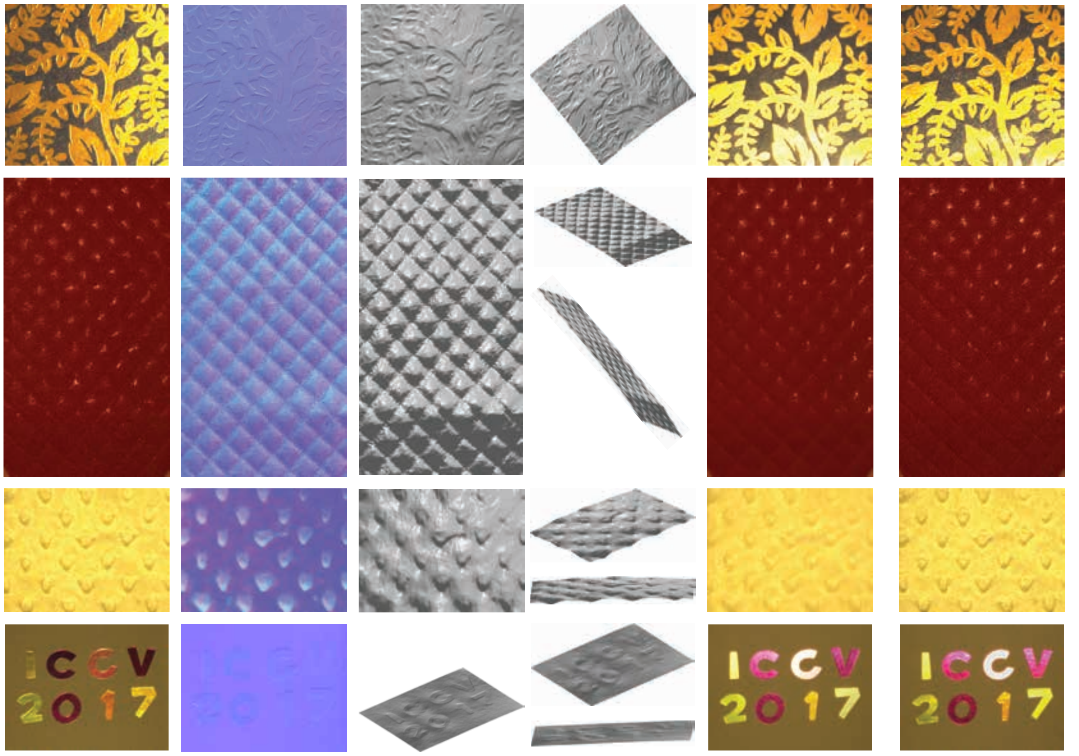
In addition to the images captured for shape and SVBRDF estimation, we capture additional images using a fixed camera and moving light source, i.e, a non-collocated setup. These “novel lighting” images are not part of the training dataset, and are used to visualize how accurately our shape and reflectance estimates generalize to directions that were not sampled. As shown in Figure 3.8, images rendered using our estimated normals and BRDFs under these novel lights closely resemble the actual captured photographs, indicating the robustness of our method.

3.4.2 Applications

Material editing. Once we reconstruct surface normals and SVBRDFs we can edit the material properties of the captured samples. This is demonstrated in Figure 3.9, where we a) swap specular and diffuse materials between two regions of the same sample, and b) transfer the specular material from one sample to a completely different sample. We re-render these edited BRDFs using the original estimated normals and view/lights. As can be seen here, our method is able to produce visually plausible results.

Material trait analysis. Previous work on recognizing material types uses specific optical setups [158] or projects raw BRDF measurements to a low-dimensional space [103]. However, these approaches are designed for objects with uniform reflectance or homogeneous BRDFs. In contrast, our technique estimates per-pixel BRDF abundances, and we can leverage this to estimate material traits at each pixel.

In order to do this, we first annotate all the materials in the MERL database with one of three unique material traits — *metal + metallic paint*, *fabric + diffuse paint* and *acrylic + plastic*. These three categories were cho-



(a) Input sample (b) Estimated normals (c) Recovered surface (d) Rendering (e) Photograph

Figure 3.8: We demonstrate shape and reflectance estimation on images captured using an iPhone 6S (a). We show the estimated normal map in false color (b) and recovered surface (c). We also compare rendered (d) results against actual captured photographs under novel lighting (e) that is not collocated with the camera.

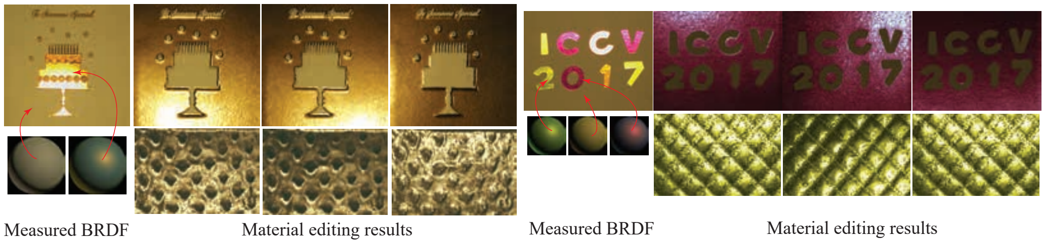


Figure 3.9: Material editing on two real samples. For the examples at the top, we compute the mean BRDF in the specular and diffuse regions of the samples (shown on the left), swap them and re-render them with the estimated normals, lights and cameras. For the examples at the bottom, we replace their SVBRDFs with the specular BRDFs from the top samples. These results are visually plausible, especially considering the fact that specular materials are likely to expose errors in geometry and material more clearly.

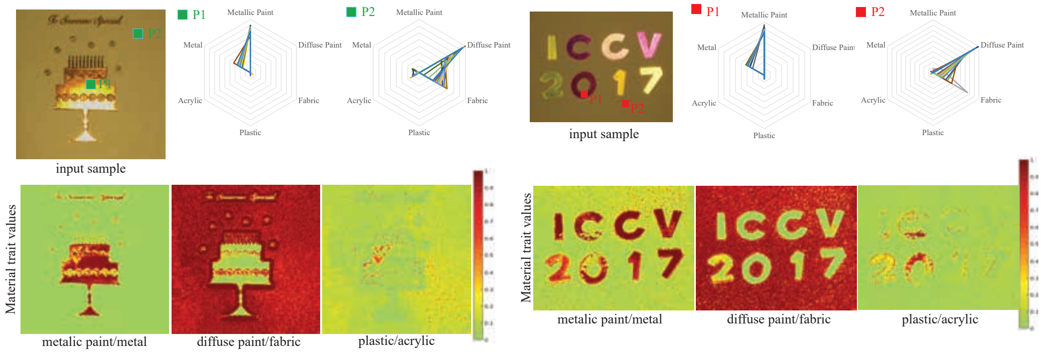


Figure 3.10: Material trait analysis on real captured data. (top) For two regions indicated by $\mathbf{p}_1$ and $\mathbf{p}_2$, we plot the associated material trait values (computed as described in Section 5.2). Pixels ($\mathbf{p}_1$) with metallic properties have large values in *metallic paint* and *metal* while pixels ($\mathbf{p}_2$) with diffuse Lambertian-like materials show large values in *diffuse paint* and *fabric*. (bottom) We visualize per-pixel material trait values for three material groups — *metallic paint+metal*, *diffuse paint+fabric*, and *plastic+acrylic*. This leads to clean, consistent material segmentations.

sen manually by visual inspection. We denote the i-th trait as $\mathcal{M}_i$. Given our abundance estimates $\widehat{\mathbf{c}}_{\mathbf{p}}$, we compute the per-pixel trait values by summing the abundances corresponding to materials with the same trait. Finally, we normalize these value so that they sum to 1:

$$\mathbf{m}_{\mathbf{p}}^{i} = \frac{\sum_{j \in \mathcal{M}_i} \widehat{\mathbf{c}}_{\mathbf{p}}(j)}{\sum_{k} \widehat{\mathbf{c}}_{\mathbf{p}}(k)}.$$

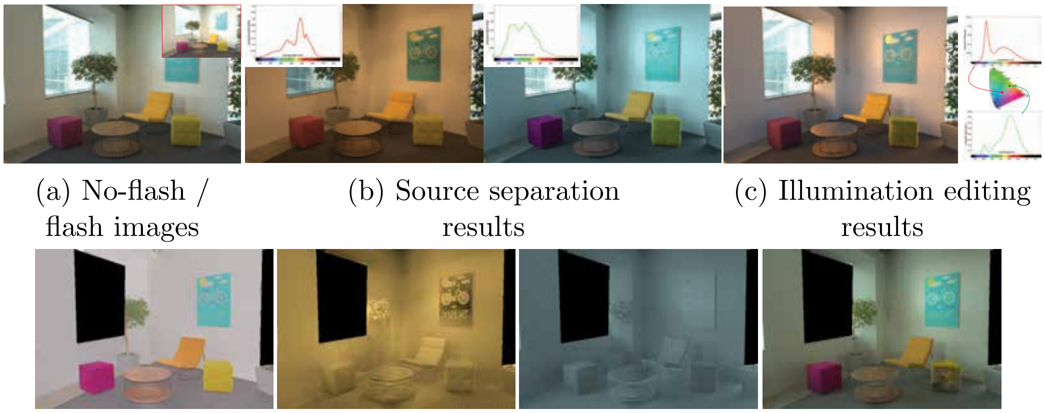
Figure 3.10 illustrates our proposed material trait analysis scheme for two datasets. Our predictions are consistent with the material properties of these samples – e.g., regions with metallic materials return high probabilities for the traits under *metal* + *metallic paint* – and accurately segment the samples into different materials.

Chapter 4

Flash photograph based illumination analysis

Real-world lighting often consists of multiple illuminants with different spectra. For example, outdoor illumination – both sunlight and skylight – differ in color temperature from indoor illuminants like incandescent, fluorescent, and LED lights. These variations in illuminant spectra manifest as color variations in captured images that are often a nuisance for vision-based analysis and photography.

In this chapter, we address the problem of explicitly separating an image into multiple images, each of which is lit by only one of the illuminants in the scene (see Figure 5.1(b)). Source separation of this form can enable a number of image editing and scene analysis applications. For example, we can change the illumination in the image by editing each illuminant image,



(a) No-flash / flash images

(b) Source separation results

(c) Illumination editing results

(d) Estimated reflectance

(e) Estimated shading for each light source

(e) Texture editing result

Figure 4.1: The scene in (a) is lit by cool sky illumination from the window on the left and warm indoor lighting from the top. Given a pair of no-flash/flash images, our method separates the no-flash image into two images lit by each of these illuminants (b) and estimates their spectral distribution (insets in (b)). Using our illuminant estimates, we are able to edit the illumination in the photograph (c) by changing the individual spectra of the light sources (insets in (c)). Given the source separation results, we also show that we are able to estimate the reflectance (d) and shading of each light source in the scene (e). This enables us to edit the texture of the scene (e) simply by operating on the reflectance.

or use the multiple images for scene analysis tasks like photometric stereo.

However, source separation is a highly ill-posed inverse problem and is especially hard from a single photograph; each pixel observation in the image combines the effect of the unknown mixture of illuminants and the unknown scene reflectance. Previous attempts at addressing these challenges either use calibrated acquisition systems [41,42] or rely on extensive user input [24–26], making it difficult to apply them at large-scale.

In this chapter, we take a step towards source separation by making use of flash photography, i.e., two photographs acquired with and without the use of the camera flash. The key insight behind our technique is that flash photography provides an image under a single illuminant, thereby enabling us to infer the reflectance spectra up to a per-pixel scale. Based on this, we derive a novel reflectance-invariant — the *Hull Constraint* — that relates light source spectra and their relative per-pixel shading to the observed intensities in the no-flash photograph. We use the Hull Constraint to separate the no-flash photograph into multiple images – each corresponding to the lighting of a unique spectra. This, in turn, enables a wide-range of capabilities including white-balancing under complex mixed illumination, the editing of the color and brightness of the separated illuminants, camera spectrum response editing and photometric stereo. The Hull constraint is independent of scene and lighting geometry; it applies equally to point and area sources as well as near and distant lighting. Figure 5.1 showcases our technique for a real-world sample.

Contributions. We propose a flash photography-based technique to analyze spatially-varying, mixed illumination. In particular, we make the following contributions:

1. We introduce a novel reflectance-invariant property of Lambertian scenes that relates illuminant spectra to observed pixel intensities.

2. We propose an algorithm to separate an image into its single-illuminant components, and present an analysis of its robustness and limitations.

3. We leverage these separated images to enable a wide variety of applications including white balancing, light editing, camera response editing, photometric stereo and intrinsic image decomposition

4.1 Related Work

In this section, we review previous works on illumination analysis as well as prior applications of flash photography.

4.1.1 Lighting analysis

Active illumination. Active illumination methods use controlled illumination to probe and infer scene properties. Controlled capture setups like a light stage [42] capture images of a person or a scene under all lighting directions and re-render photorealistic images under arbitrary illumination [46,118]. Another class of techniques rely on projector-camera systems to probe and separate light transport in a scene [106]. While active techniques can enable high-quality illumination analysis and editing, these systems are complex and expensive. In contrast, we propose a simple capture process that uses a camera flash, available on most cameras and mobile devices, to enable a number of illumination analysis, editing and reconstruction tasks.

Passive illumination. Passive illumination methods aim to estimate scene properties from images captured as-is under natural illumination. Barron and Malik [15,17] estimate shape, reflectance, and illumination for a single object captured under low-frequency distant lighting. Johnson and Malik [87] use spectral variations in real-world illumination to recover shape from shading information. Both methods rely on scene priors that are often violated on real-world scenes with complex geometry, reflectance, and spatially-varying lighting. In contrast, we demonstrate that the use of flash photography can lead to high-quality lighting (and shape) estimates without the same restrictive assumptions. Recently, deep learning-based methods have been proposed to infer illumination [59, 74] from a single image. However, these methods do not support pixel-level image editing, which our method does by explicitly separating an input image into its constituent components.

Color constancy. Color constancy — the problem of correcting for the illuminant spectrum — is a closely related light analysis problem, and has been extensively studied [61]. Previous work models the effect of changing the illumination spectral distribution as a (typically linear) transformation of the observed pixel intensities. The seminal work of Finlayson et al. [52, 55] demonstrates real-world reflectance and illumination spectra lie in low-dimensional spaces, allowing for the use of a diagonal transformation. Chong et al. [38] build on this to derive conditions for the basis that can "best" support diagonal color constancy. Current color constancy methods range from physics/low-level feature-based methods [44, 62, 147] to learning-based approaches [14, 31, 88] to user-driven interactive solutions [26, 77]. The vast majority of these methods assume a single illuminant in the scene. While our approach is built on top of diagonal color constancy techniques, we can handle multiple illuminants and can go beyond color constancy and separate the captured image into constituent images lit by individual illuminants.

4.1.2 Flash photography

Flash photography refers to techniques that capture two images of a scene — with and without flash illumination. It has been used for image denoising [48, 119], deblurring [162], artifact removal [48], non-photorealistic rendering [122], foreground segmentation [142] and matting [143]. More recently Hui et al. [80] propose a flash photography-based white balancing method for mixed illumination. However, the techniques in this chapter are derived from a physically-accurate image formation model and are based on a novel reflectance-invariant, the Hull constraint, which enables explicit separation of the contribution of different light sources at each pixel. Our analysis enables a number of applications that are not possible with prior work [80], including light editing, and two-shot photometric stereo.

4.1.3 Intrinsic image

Intrinsic image algorithms [18] that separate images into reflectance and shading components often assume that there is a single (usually white) illuminant and the induced shading is spatially smooth [20, 66, 90, 135, 159]. While these techniques work for some scenarios, real world scenes often have complex geometry and multiple spatially-varying light sources that cannot be model using simple priors. To account for that, Barron et al. [15, 17] extend the idea by incorporating the illumination into the optimization framework. Given rich training samples of the natural images, Bell et al. [20] address the problem by utilizing Conditional Random Field (CRF) to characterize the relationship between reflectance and illumination in the scene. More recently, deep neural network-based techniques have been widely adopted for estimating reflectance and illumination from a single image [95,97]. However, these techniques focus on the scene with single illumination. In contrast, the scenes we focus on have significantly more complex illumination variations.

4.2 The Hull Constraint

Given an image of a scene lit by a mixture of illuminants — the *no-flash* image — our goal is to estimate the contribution of each illuminant to the observed pixel intensities. In this section, we set up the image formation model and derive a novel constraint between the observed no-flash/flash pixel intensities and the contributions of each scene illuminant to the scene appearance.

4.2.1 Problem setup and image formation

We assume that the scene is Lambertian and is imaged by a three-channel color camera. The intensity of the no-flash image observed at a pixel $\mathbf{p}$ in the k-the color channel ($k \in \{r, g, b\}$) is

$$I^k_{\text{nf}}(\mathbf{p}) = \int_\lambda \rho_{\mathbf{p}}(\lambda) S^k(\lambda) \ell_{\mathbf{p}}(\lambda) d\lambda, \tag{4.1}$$

where $\rho_{\mathbf{p}}$ is the reflectance spectra, S^k is the camera spectral response for the k-th channel and $\ell_{\mathbf{p}}(\lambda)$ is the light spectra at pixel $\mathbf{p}$. When the scene is lit by N light sources, the light spectra at pixel $\mathbf{p}$ can be expressed as

$$\ell_{\mathbf{p}}(\lambda) = \sum_{i=1}^{N} \eta_i(\mathbf{p}) \ell_i(\lambda),$$

where $\ell_i(\lambda)$ is the spectra of the i-th light source and $\eta_i(\mathbf{p})$ is the shading corresponding to the i-th source at pixel $\mathbf{p}$. The shading term $\eta_i(\mathbf{p})$ is assumed to be non-negative. Note that, by not modeling $\eta_i(\mathbf{p})$ with an analytical expression, we can accommodate point, extended and area light sources. Since the illumination spectra $\{\ell_1, \ldots, \ell_N\}$ are not pixel dependent, any spatial light fall-off is captured in the shading term. With this, (4.1) can be written as

$$I_{\mathrm{nf}}^k(\mathbf{p}) = \int_\lambda \rho_{\mathbf{p}}(\lambda) S^k(\lambda) \left(\sum_{i=1}^{N} \eta_i(\mathbf{p}) \ell_i(\lambda) \right) d\lambda. \tag{4.2}$$

Estimating the reflectance, shading and illumination parameters as well as separating the no-flash photograph into N photographs — one for each of the N light sources — are hard inverse problems. The parameters of interest, namely $\rho_{\mathbf{p}}$ and ℓ_i, are infinite-dimensional. Further, the multi-linear encoding of the reflectance, shading and illumination parameters in the image intensities leads to a highly-ambiguous solution space. To resolve these challenges, we make two key assumptions.

Assumption 1 — Reflectance and illumination subspaces. We assume that the reflectance and illumination spectra in the scene are well-approximated by low-dimensional subspaces. Given a reflectance basis $B_R(\lambda) = [\widetilde{\rho}_1(\lambda) \ldots \widetilde{\rho}_{M_1}(\lambda)]$ and an illumination basis $B_L(\lambda) = \left[\widetilde{\ell}_1(\lambda) \ldots \widetilde{\ell}_{M_2}(\lambda)\right]$, we can write

$$\rho_{\mathbf{p}}(\lambda) = B_R(\lambda)\ \mathbf{a}_{\mathbf{p}}, \quad \ell_i(\lambda) = B_L(\lambda)\ \mathbf{b}_i.$$

Here, $\mathbf{a}_{\mathbf{p}} \in \mathbb{R}^{M_1}$ are the reflectance coefficients at pixel $\mathbf{p}$ and $\mathbf{b}_i \in \mathbb{R}^{M_2}$ are the illumination coefficients for the i-th source. To resolve the ambiguity in the definition of the shading, we assume that the lighting coefficients are unit-norm, i.e., $\|\mathbf{b}_i\|_2 = 1$; hence, the illumination coefficients are points on the 2D sphere. Given this, we write (5.1) as:

$$I^k_{\text{nf}}(\mathbf{p}) = \mathbf{a}_{\mathbf{p}}^\top E^k \sum_{i=1}^{N} \eta_i(\mathbf{p})\mathbf{b}_i. \tag{4.3}$$

Here, E^k is the $M_1 \times M_2$ matrix defined as

$$E^k(i,j) = \int_\lambda \widetilde{\rho}_i(\lambda) S^k(\lambda) \widetilde{\ell}_j(\lambda) d\lambda,$$

and can be precomputed from a database of reflectance and illumination spectra. Finally, as a consequence of having 3-color images, we will need to restrict $M_1 = M_2 = 3$. Real world reflectance and illumination spectra are known to be well-approximated by low-dimensional subspace — an insight that is used extensively in the color constancy [38,52,55,58]. We will discuss additional details on the choice of basis in Section 4.5.

Assumption 2 — Availability of a flash photograph. We resolve the multi-linearity of the unknown parameters by having access to a flash photograph of the scene. In the flash image I_{f}, the intensity observed at pixel $\mathbf{p}$ is given by:

$$I^k_{\text{f}}(\mathbf{p}) = I^k_{\text{nf}}(\mathbf{p}) + \int_\lambda \rho_{\mathbf{p}}(\lambda) S^k(\lambda) \eta_{\text{f}}(\mathbf{p}) \ell_{\text{f}}(\lambda) d\lambda, \tag{4.4}$$

where $\eta_{\text{f}}(\mathbf{p})$ denotes the shading at $\mathbf{p}$ induced by the flash, and the spectra of the flash ℓ_{f} is assumed to be known via a calibration process. Further, under the reflectance and illumination subspace modeling above, we can write

$$I_{\text{f}}^{k}(\mathbf{p}) = I_{\text{nf}}^{k}(\mathbf{p}) + \mathbf{a}_{\mathbf{p}}^{\top} E^{k} \eta_{\text{f}}(\mathbf{p})\mathbf{f}, \tag{4.5}$$

where $\mathbf{f}$ denotes the illumination coefficients for the flash spectra. We now derive a novel constraint that encodes both the illuminant spectra as well as their shadings at each pixel.

4.2.2 The Hull Constraint

The centerpiece of our approach is a novel reflectance-invariant condition that we call the Hull Constraint. The hull constraint is derived by performing the following three operations (see Figure 4.2 for a visual guide).

Step 1 — Estimate the pure flash image. The pure-flash image I_{pf} is obtained by subtracting the no-flash image from the flash image:

$$I_{\text{pf}}^{k}(\mathbf{p}) = I_{\text{f}}^{k}(\mathbf{p}) - I_{\text{nf}}^{k}(\mathbf{p}) = \mathbf{a}_{\mathbf{p}}^{\top} E^{k} \eta_{\text{f}}(\mathbf{p})\mathbf{f}. \tag{4.6}$$

Step 2 — Solve for reflectance coefficients. We now have 3 intensity measurements — one per color channel — at $\mathbf{p}$, and 3 unknowns for $\boldsymbol{\alpha}_{\mathbf{p}} = \eta_{\text{f}}(\mathbf{p})\mathbf{a}_{\mathbf{p}}$. This enables us to solve for $\boldsymbol{\alpha}_{\mathbf{p}}$, which corresponds to the reflectance coefficients up to a per-pixel scale $\eta_{\text{f}}(\mathbf{p})$.

Step 3 — Estimate $\Gamma(\mathbf{p})$. Since $\frac{\boldsymbol{\alpha}_\mathbf{p}}{\|\boldsymbol{\alpha}_\mathbf{p}\|} = \frac{\mathbf{a}_\mathbf{p}}{\|\mathbf{a}_\mathbf{p}\|}$, we can substitute $\boldsymbol{\alpha}$ to express (4.3) as:

$$I^k_{\text{nf}}(\mathbf{p}) = \|\mathbf{a}_\mathbf{p}\| \left(\frac{\boldsymbol{\alpha}^T_\mathbf{p}}{\|\boldsymbol{\alpha}_\mathbf{p}\|} \right) E^k \sum_{i=1}^{N} \eta_i(\mathbf{p})\mathbf{b}_i. \tag{4.7}$$

As before, we are able to solve for $\boldsymbol{\beta}(\mathbf{p})$, defined as

$$\boldsymbol{\beta}(\mathbf{p}) = \|\mathbf{a}_\mathbf{p}\| \sum_{i=1}^{N} \eta_i(\mathbf{p})\mathbf{b}_i. \tag{4.8}$$

Normalizing $\boldsymbol{\beta}(\mathbf{p})$ gives us $\Gamma(\mathbf{p}) = \boldsymbol{\beta}(\mathbf{p})/\|\boldsymbol{\beta}(\mathbf{p})\|$ that is invariant to the reflectance. We can now state the Hull constraint, which is the main contribution of this chapter.

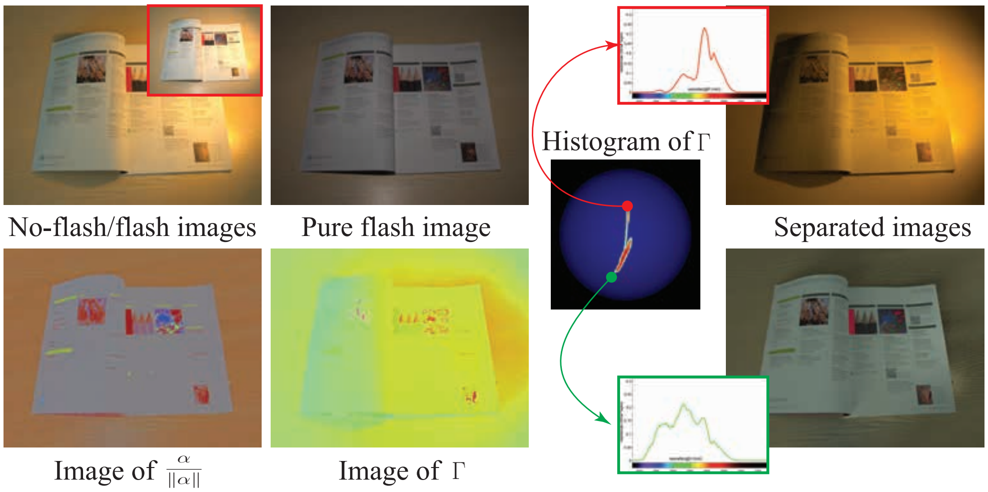
Proposition 1 (The Hull Constraint)**.** *The term* $\Gamma(\mathbf{p})$ *lies in the conic hull of the coefficients* $\{\mathbf{b}_1, \ldots, \mathbf{b}_N\}$*, i.e.,*

$$\boxed{\Gamma(\mathbf{p}) = \frac{\boldsymbol{\beta}(\mathbf{p})}{\|\boldsymbol{\beta}(\mathbf{p})\|} = \sum_{i=1}^{N} z_i(\mathbf{p})\mathbf{b}_i, \quad z_i(\mathbf{p}) \geq 0.} \tag{4.9}$$

The relative shading term $z_i(\mathbf{p})$ is defined as

$$z_i(\mathbf{p}) = \frac{\eta_i(\mathbf{p})}{\|\sum_j \eta_j(\mathbf{p})\mathbf{b}_j\|}. \tag{4.10}$$

This term captures the fraction of the shading at a scene pixel that comes from one light source, relative to all the light sources, hence the term *relative*



No-flash/flash images

Pure flash image

Separated images

Image of $\frac{\alpha}{||\alpha||}$

Image of Γ

Figure 4.2: Visualization of our processing pipeline. From the input image pair, we compute the pure flash image as well as values of the $\boldsymbol{\alpha}$ and Γ at each pixel. We visualize $\boldsymbol{\alpha}/\|\boldsymbol{\alpha}\|$ and Γ as 3-color images by integrating them with the reflectance and illumination bases, respectively, and the camera spectral response. Note that $\boldsymbol{\alpha}/\|\boldsymbol{\alpha}\|$ encodes the scene's reflectance while Γ, being reflectance-invariant, encodes the shading and illumination. The histogram of Γ over the sphere provides an estimate of the illumination spectra as well as the separated images.

shading. Further, $\Gamma(\mathbf{p})$ belongs to $\mathbb{S}^2$ space since it is unit-norm.

The key insight of the Hull constraint is that $\Gamma(\mathbf{p})$, a quantity that can be estimated from the no-flash/flash image pair, provides an encoding of the illumination coefficients as well as the relative shading. We can hence derive these parameters as well as perform source separation by studying properties of $\Gamma(\mathbf{p})$ over the entire image.

4.3 Source Separation with the Hull Constraint

Recall, from Proposition 1, that $\Gamma(\mathbf{p})$ lies in the conic hull formed by the lighting coefficients $\{\mathbf{b}_1, \ldots, \mathbf{b}_N\}$. We now describe methods to estimate the illuminant spectrum as well as perform source separation from the set $\mathcal{G} = \{\Gamma(\mathbf{p});\ \forall \mathbf{p}\}$. Our methods rely on fitting the tightest conic hull to the set $\mathcal{G}$ and identifying the corners of the estimated hull. Additionally, we derive sufficient/necessary conditions when the resulting estimates are meaningful. We begin by discussing the conditions for the identifiability of a light source.

4.3.1 Identifiability of a light source

We observe that a light source is identifiable only if its coefficients lie outside the conic hull of the coefficients of the remaining light sources. If this were not the case, then its contribution to a scene point can be explained by the remaining lights. Hence, only light sources whose coefficients lie at corners of the conic hull of $\{\mathbf{b}_1, \ldots, \mathbf{b}_N\}$ are identifiable given the flash/no-flash image pair. Without any loss in generality, we assume that all light sources are identifiable. Therefore, if we can identify the conic hull of the light sources $\mathcal{L} = \text{conic-hull}\{\mathbf{b}_1, \ldots, \mathbf{b}_N\}$, we can estimate the light source coefficients as the corner points of this set. While we do not have an a-priori estimate of $\mathcal{L}$, we can estimate it from the set $\mathcal{G} = \{\Gamma(\mathbf{p});\ \forall \mathbf{p}\}$. Recall that, from Proposition 1, $\mathcal{G} \subseteq \mathcal{L}$. We next explore sufficient conditions under which the conic hull of $\mathcal{G}$ is equal to $\mathcal{L}$; when this happens, we can estimate the light source coefficients as the corner points of the conic hull of $\mathcal{G}$.

Proposition 2 (Presence of "pure" pixels)**.** *Under ideal imaging conditions (absence of noise, non-Lambertian surfaces, etc.), the conic hull of* $\mathcal{G}$ *is equal to* $\mathcal{L}$ *if, for each light source, there exists a pixel that is purely illuminated by that light source, or, equivalently,*

$$\forall \mathbf{b} \in \{\mathbf{b}_1, \ldots, \mathbf{b}_N\},\ \exists\ \Gamma(\mathbf{p}') = \mathbf{b}.$$

When there are pure pixels for each light source, then the set $\mathcal{G}$ will include the illuminant coefficients which are also the corners of the conic hull $\mathcal{L}$. Therefore, the conic hull of $\mathcal{G}$ will be identical to $\mathcal{L}$. Note that pure pixels can be found in shadow regions since shadows indicate the absence of light source(s). The pure pixel assumption is thus satisfied when the scene geometries are sufficiently complex to exhibit a wide array of cast and attached shadows. The more complex the scene geometry, the more likely it is that we satisfy the condition in Proposition 2.

In addition to pure pixels or corners, we can also fit the hull by identifying its edges. Edges of the cone correspond to points that are in the shadow of all but two sources. As with pure pixels, shadows play a pivotal role in recovering the hull from its edges.

4.3.2 Estimating illuminant coefficients

Given the set $\mathcal{G}$, the number of identifiable light sources is simply the number of corners in the tightest conic hull. Hence, we expect the set $\mathcal{G}$ to be concentrated about a point when there is a single light source, an arc with two sources, and so on (see Figure 4.3). We can use specialized techniques to estimate the parameters in each case (see detailed pseudo-code in the supplemental material).

- $N = 1$ — While not particularly interesting in the context of source separation, we use the robust mean of $\mathcal{G}$ as the coefficients of the single light source.

- $N = 2$ — We use RANSAC to robustly estimate the arc on $\mathbb{S}^2$ with maximum inliers. The end points of this arc are associated with the illuminant coefficients; this estimate will correspond to the true coefficients if there were “pure pixels” in the no-flash photograph for each of the light sources.

- $N = 3$ — We project the set $\mathcal{G}$ onto the tangent plane at its centroid and fit the triangle with least area onto the projected points. Fitting polyhedra onto planar points has been extensively studied in computational geometry [12,49,104,110,117]. We use the method in Parvu et al. [117] to determine the triangle and the associated vertices.

- $N \geq 4$ — The procedure used for three light sources can potentially be applied to higher number of sources. However, as we will see next, even if

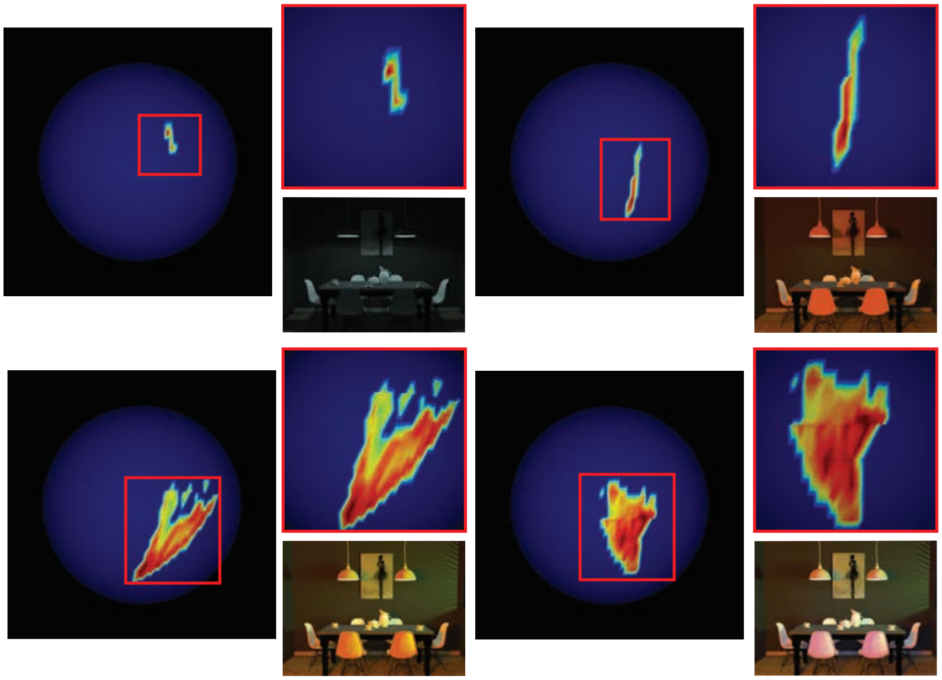


Figure 4.3: Visualization of $\mathcal{G}$ as a histogram on $\mathbb{S}^2$ for different numbers of sources. The histogram takes progressively complex shapes as the number of sources increase (from 1, top left, to 4, bottom right).

we can estimate the lighting coefficients, source separation with a three-color camera cannot be performed when $N \geq 4$.

For the results in the chapter, we manually specify the number of light sources (typically, 2 or 3) and use the corresponding algorithm to extract the corners. Given the estimated lighting coefficients $\{\widehat{\mathbf{b}}_1, \ldots, \widehat{\mathbf{b}}_N\}$, we can estimate the relative shading at each pixel.

4.3.3 Estimating the relative shading

Given $\Gamma(\mathbf{p})$ and estimates of the lighting coefficients $\{\widehat{\mathbf{b}}_1, \ldots, \widehat{\mathbf{b}}_N\}$, we simply solve the linear equations in (4.9) under non-negativity constraints to estimate the relative shading $\{z_i(\mathbf{p}), i = 1, \ldots, N\}$. It is easily shown that there is a unique solution when $\Gamma(\mathbf{p}) \in \text{conic-hull}\{\widehat{\mathbf{b}}_1, \ldots, \widehat{\mathbf{b}}_N\}$ and $N \leq 3$ (see supplemental material). When $N > 3$, we can obtain multiple solutions to the relative shading — a limitation that stems from using 3-color cameras.

4.3.4 Lighting separation

Once we have the illumination coefficients $\{\widehat{\mathbf{b}}_1, \ldots, \widehat{\mathbf{b}}_N\}$ and the relative shading $\{\widehat{z}_i(\mathbf{p})\}$, we can separate the no-flash photograph into N photographs. Specifically, for the j-th light source, we would like to estimate

$$I^k_{\text{sep},j} = \mathbf{a}_{\mathbf{p}}^\top E^k \eta_j(\mathbf{p}) \mathbf{b}_j.$$

An estimate of this image is obtained as

$$\widehat{I}^k_{\text{sep},j}(\mathbf{p}) = \|\boldsymbol{\beta}(\mathbf{p})\| \boldsymbol{\alpha}_{\mathbf{p}}^\top E^k \widehat{z}_j(\mathbf{p}) \widehat{\mathbf{b}}_j. \tag{4.11}$$

4.4 Reflectance estimation

While our approach provides the estimates for the reflectance spectra up to a per-pixel scale, it is desirable to obtain the reflectance of the scene without such ambiguities. In particular, given all the separated images in (4.18), our goal here is to estimate the $\|\mathbf{a}\|$ at each pixel. Recall, from (4.15), we can write the separated images as

$$\widehat{I}^k_{\text{sep},j}(\mathbf{p}) = \frac{\boldsymbol{\alpha}_{\mathbf{p}}^\top}{\|\boldsymbol{\alpha}_{\mathbf{p}}\|} E^k \widehat{\mathbf{b}}_j \eta_j(\mathbf{p}) \|\mathbf{a}_{\mathbf{p}}\|. \tag{4.12}$$

We are able to solve for $\eta_j(\mathbf{p})\|\mathbf{a}_{\mathbf{p}}\|$ and denote the term as

$$q_j(\mathbf{p}) = \eta_j(\mathbf{p})\|\mathbf{a}_{\mathbf{p}}\|$$

Similarly, we can also do this for the pure flash image and get the results as

$$q_{\text{pf}}(\mathbf{p}) = \eta_f(\mathbf{p})\|\mathbf{a}_{\mathbf{p}}\|$$

Given these three (scene with two lights) or four (scene with three lights) resulting images, it requires the separation of them into a product of the norm of the reflectance coefficients and illuminant induced shadings. The problem here is similar in spirit to intrinsic decomposition techniques that seek to separate an image into a per-pixel product of reflectance, which is characterized by the material property, and illumination layer, which is characterized by the light color and the scene geometry. While significant progress has been made on the intrinsic decomposition problem, the vast majority of the state-of-the-art works rely on additional priors to disambiguate reflectance and illumination spectrum from the scene content, such as incorporating depth maps [], non-local texture priors [20, 135] and gradients properties of reflectance and shading [151]. However, these approaches highly depend on the scene-specific tuned parameters and always require sophisticated optimization scheme.

Our approach differs from these techniques in that we are able to estimate the illuminant spectrum in the scene, which significantly constrict the underlying solution space. In particular, our algorithm takes the separated images together with pure flash images as the input. Note that we "white balance" for all the above images, making the color of illuminant neutral. That is, we can write the white balanced separated images as

$$\widehat{I}^k_{\text{sep},j}(\mathbf{p}) = \frac{\boldsymbol{\alpha}_{\mathbf{p}}^\top}{\|\boldsymbol{\alpha}_{\mathbf{p}}\|} E^k \widehat{l} \eta_j(\mathbf{p}) \|\mathbf{a}_{\mathbf{p}}\|. \tag{4.13}$$

We assume that the gradient of reflectance norm $\|\mathbf{a}_{\mathbf{p}}\|$ tends to be large while the illuminant induced shadings $\eta_i(\mathbf{p})$ are varied smoothly, making their gradients relatively small. We do this by formulating the problem as follows

$$I^k_{\text{wb},j}(\mathbf{p}) = \mathbf{a}_{\mathbf{p}}^\top E^k \eta_j(\mathbf{p}) \mathbf{b}_{\text{wb}}, \tag{4.14}$$

where $\mathbf{b}_{\text{wb}}$ denotes the illumination coefficients for the white light.

Once we have the input images, we propose a coarse-to-fine scheme to reconstruct the reflectance. This enables the capability in the texture editing for the scene under the mixture of illumination as we will show later in the chapter. In particular, given the resulting images $[q_1(\mathbf{p}), q_2(\mathbf{p}), \ldots, q_{\text{pf}}(\mathbf{p})]$, we solve for the norm of the reflectance $\|\mathbf{a_p}\|$ and the induced shading $\eta_i(\mathbf{p})$. Similar to previous work in the field, we work in the log-domain to transform the product into the sum $\log q_i(\mathbf{p}) = \log \eta_i(\mathbf{p}) + \log \|\mathbf{a_p}\|$.

Reflectance and shading estimation. With the notations and formulation, we can formulate the problem by solving a least square energy term defined as

$$E(r, s) = \sum_{i=1}^{N} \|q_i - s_i - r\|_2^2,$$

where r denotes the $\|\mathbf{a_p}\|$ and s_i denotes the shading $\eta_i(\mathbf{p})$. We incorporate the ℓ_1 prior on the gradient of the reflectance, by assuming that the reflectance are piece-wise linear, such that the scene under consideration can be generated from a small number of unique reference reflectance functions. In the contrast, we incorporate the ℓ_2 prior on the gradients of the shading, by assuming that the illumination in the scene is smoothly distributed. We put all terms together to formulate the problem as

$$\widehat{s}_i,\ \widehat{r} = \arg\min_{s_i, r} \sum_{i=1}^{N} \|q_i - s_i - r\|_2^2 + \lambda_r \|\nabla r\|_1 + \sum_{i=1}^{N} \lambda_s \|\nabla s_i\|_2.$$

Estimating shading. Given the estimate of (k)-th iteration, we perform the gradient descent to the above problem to obtain the estimate at $(k+1)$-th iteration. In particular, we use the difference matrix D to characterize the gradient operation as

$$\widehat{s}_i,\ \widehat{r} = \arg\min_{s_i, r} \sum_{i=1}^{N} \|q_i - s_i - r\|_2^2 + \lambda_r \|Dr\|_1 + \sum_{i=1}^{N} \lambda_s \|Ds_i\|_2.$$

We first compute the gradient for the smooth data fidelity term and update the s_i as

$$\widehat{s}_i^{(k+1)} = \widehat{s}_i^{(k)} + 2t(\widehat{s}_i^{(k)} + \widehat{r}^{(k)} - q_i + \lambda_s D^\top D\widehat{s}_i^{(k)}),$$

where t is the updating step.

Estimating Reflectance. To estimate the reflectance, we denote $Dr = z$ and incorporate the alternating direction method of multipliers to update the shading term.

$$\widehat{r}, \widehat{z}, \widehat{p}, \widehat{\mu} = \arg\min_{r,z,\mu} \sum_{i=1}^{N} \|q_i - s_i - r\|_2^2 + \lambda_r \|z\|_1 + \frac{p}{2}\|Dr - z + \mu\|_2^2.$$

The estimates can be obtained by

$$\widehat{r}^{(k+1)} = \arg\min_{r} \sum_{i=1}^{N} \|q_i - s_i - r\|_2^2 + \frac{p}{2}\|Dr - \widehat{z}^{(k)} + \widehat{\mu}^{(k)}\|_2^2,$$

$$\widehat{z}^{(k+1)} = \arg\min_{z} \lambda_r \|z\|_1 + \frac{p}{2}\|D\widehat{r}^{(k)} - z + \widehat{\mu}^{(k)}\|_2^2,$$

$$\widehat{\mu}^{(k+1)} = \widehat{\mu}^{(k)} + D\widehat{r}^{(k)} - \widehat{z}^{(k)}.$$

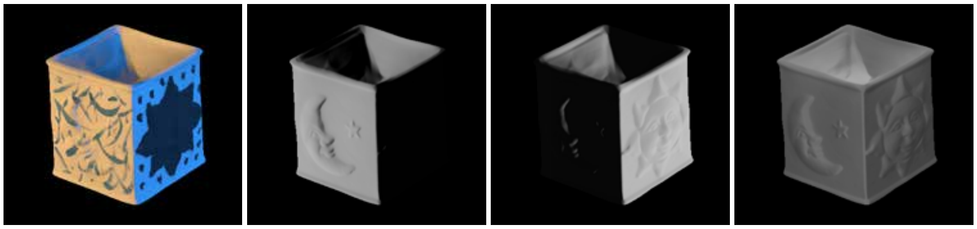
Given the estimates for r and s_i, we reconstruct the $\|\mathbf{a_p}\|$ and η_i for each separated image. We further incorporate the estimated relative shadings to better constrain the estimation of the reflectance. Specifically, given the resulting images $[q_1(\mathbf{p}), q_2(\mathbf{p}), \ldots, q_{\text{pf}}(\mathbf{p})]$, we are able to derive the relative shadings by dividing each image by the sum of all the images as

$$\widetilde{q_i} = \frac{q_i}{\sum_{j=1}^{N} q_j} = \frac{\eta_i}{\sum_{j=1}^{N} \eta_j}.$$

An example is shown in Figure 4.4. To incorporate the relative shadings, we require the gradient of the estimated reflectance to be small for the regions where the relative shadings have large gradient values. This amounts to thresholding the image gradient according to the relative shadings. Specifically, we have

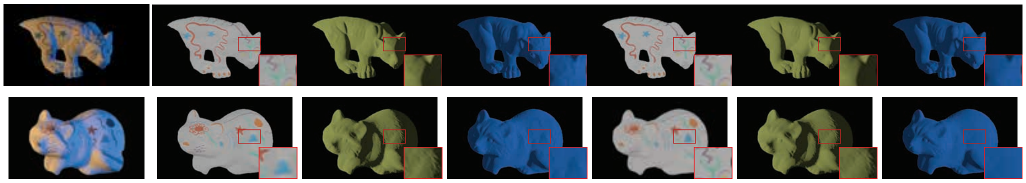
$$\nabla \widehat{r}^{(k)} = \begin{cases} \nabla \widehat{r}^{(k)} & \text{if } \nabla \widetilde{q_i} < T_s \ \forall i \\ 0 \quad \text{otherwise} \end{cases}$$

In this chapter, we set T_s to be 0.2 for all the results. Given the estimate of $\|\mathbf{a_p}\|$, we can compute for the shading η_i by dividing $\|\mathbf{a_p}\|$ from the q_i. We showcase the difference by incorporating the shading prior in Figure 4.5. As can be seen, the regions with high-frequency texture can be correctly removed from shadings by incorporating the priors, while these are transfered into shadings in the initial reflectance estimation.



(a) Input (b) Relative shadings $\widetilde{q}_i$

Figure 4.4: Visualization of the relative shadings for two-light source in a synthetic scene (a). We show the relative shadings for both separated images and pure flash image. (b)



(a) No-flash image (b) Initial reflectance and shading estimates (c) Incorporating shading prior

Figure 4.5: We show the improvement by incorporating the shading prior the for synthetic results on two-light source (a). Processing with initial reflectance and shading algorithm, it appears incorrectly in separated shadings (as shown in the insets) (b). By incorporating the shading priors, we can see that it does not appear in the shading estimates (c).

4.5 Evaluation and Applications

We characterize the performance of the proposed methods by evaluating light separation and showcasing its potential in a number of applications.

Capture setup for real data. The flash/no-flash images were captured using a Nikon D800 and a Speedlight SB-800 flash, with the camera mounted on a tripod and operated under aperture-priority mode. The images were captured in raw format and demosaiced under a linear response using DCRaw [1].

Finally, the flash spectrum was assumed to be flat, i.e., $\ell_f(\lambda)$ in (4.4) was assumed to be a constant.

Selection of reflectance and illumination bases. We used the measured database for reflectance [69] and illumination [2] to learn two three-dimensional subspaces, one each for reflectance and illumination. All the results in this chapter were obtained with the same pair of bases, which we learned using a weighted PCA model, with the camera spectral response providing the weights. We observed that this technique outperformed an unweighted PCA as well as the joint learning of subspaces [38]. The supplemental material provides a detailed evaluation on synthetic scene with comparisons to alternate strategies.

Pruning $\mathcal{G}$. To reduce effects of measurement noise and model mismatch, we build a histogram of $\mathcal{G}$ by dividing the sphere into 100×100 bins and counting the occurrence of $\Gamma(\mathbf{p})$ in each bin. We remove points in sparsely populated regions; typically, points in bins that have less than 100 pixels are removed from $\mathcal{G}$.

4.5.1 Evaluation of lighting separation

We report the performance of our source separation technique on both synthetic and a wide-range of real-world scenes.

Synthetic experiments. We evaluate the source separation technique on realistically-rendered scenes using the MITSUBA rendering engine [85]. Specifically, when we simulate the scene, we select two light spectra from [2] and compute the errors for both separated images, as well as the light coefficients, against the ground truth. For the scene with two lights, we report these errors as a function of varying angular difference for the ground truth spectra in Figure 4.6. We observe that the SNR values of the source separation are larger than 30dB for most of the lighting spectra, even for the worst case, i.e. (1° in the angular difference), the SNR value can still be achieved at 16dB, suggesting the robustness of our technique. We also showcase the angular error against the ground truth coefficients. Note that the angular error increases with the difference between the lighting spectra. This is due to the fact that there is a decrease of the conic hull characterized by Γ as we moved one lighting spectra away from the other. In particular, the potential region characterized the estimation errors also shrinks with the increase similarity in lighting spectra.

Similarly, for the scene with three lights, we select the measured lighting spectra from the dataset [2] and make sure the smallest angular difference of the illumination coefficients for these selected lighting spectra is larger than 20° in degrees. In Figure 4.7, we report the errors against the ground truth for 26 rendered flash/no-flash image pairs. As can be seen, our algorithm is able to return the results larger than 20 dB for all the datasets. We also include the visual results for two selected samples in Figure 4.7. As can

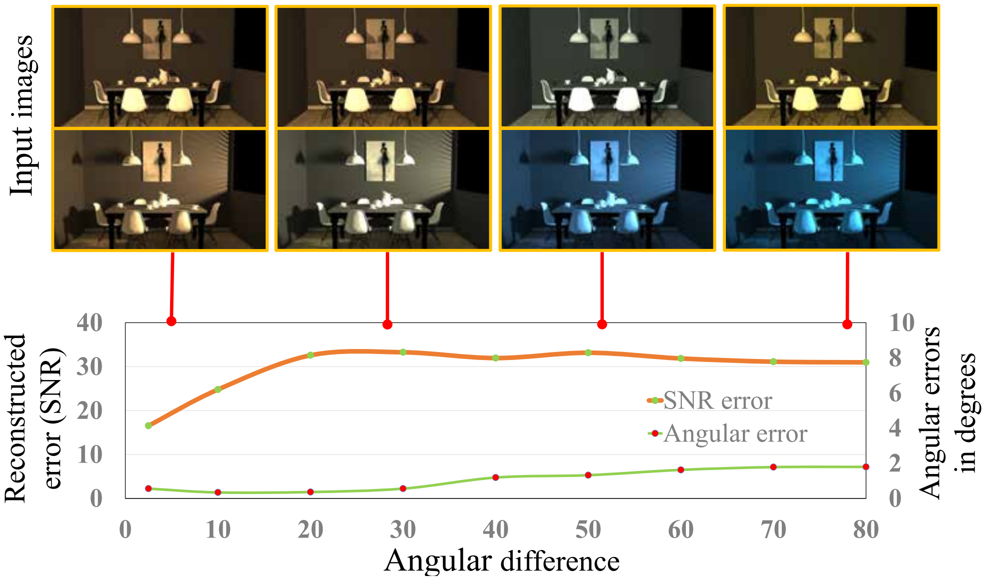


Figure 4.6: Evaluation of the two-light source separation on synthetic dataset. For each scene (source credit to [21]), we pick two measured illuminant spectra from LSPDD database. We plot the errors measured against the ground truth constituent images as a function of angular difference between ground truth lighting coefficients.

be seen, our algorithm is able to capture the color and shadings for each illuminant as well as produce the results very close to the ground truth.

Scenes with two lights. In Figure 4.8, we demonstrate our technique on the scene with two lights sources and compare with ground truth captures. Ground truth photographs were obtained by turning off the indoor light sources to obtain the outdoor illuminated scene and then subtracting this from the no-flash image to obtain the photograph with respect to the indoor illumination. We also compare against a simple non-negative matrix factorization (NNMF) as well as the technique proposed in Hsu et al. [77]. Naively

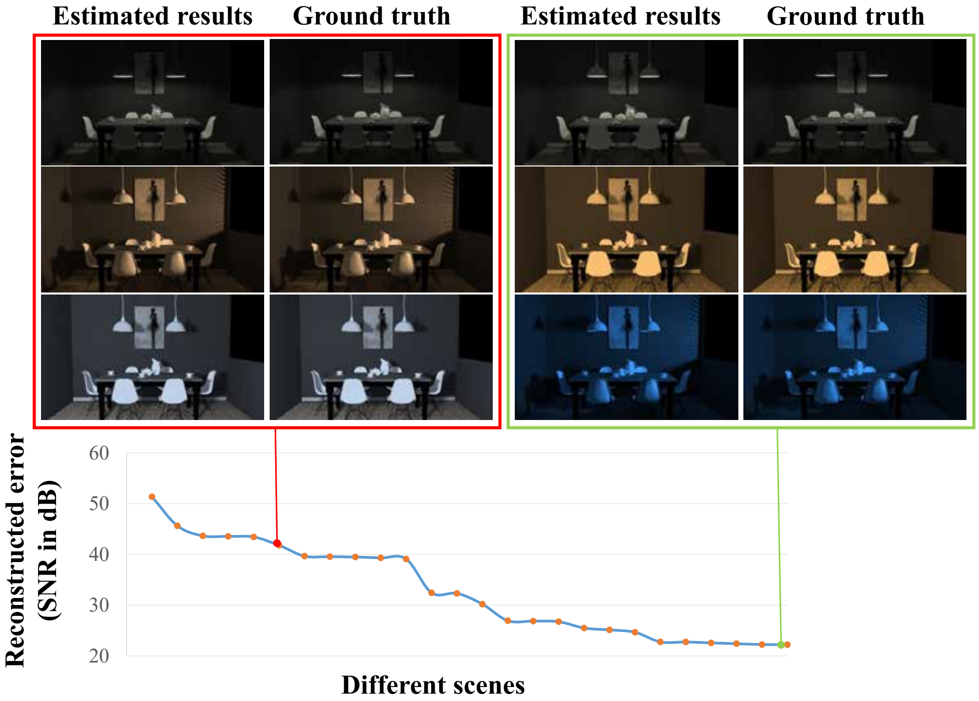
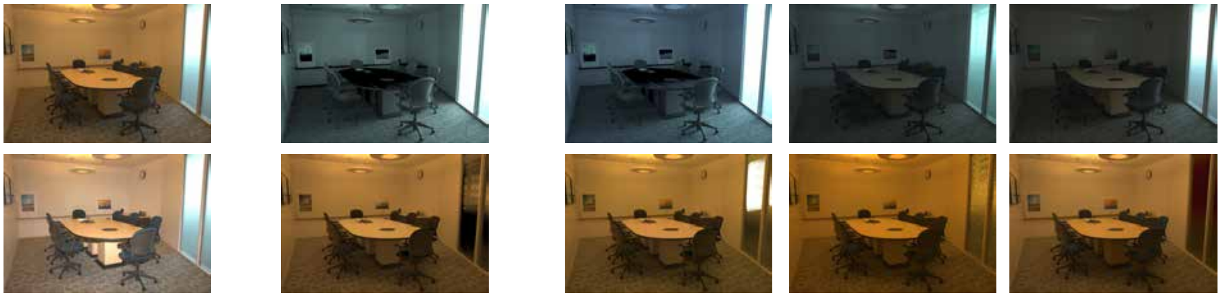


Figure 4.7: Evaluation of the three-light source separation. For each generated scene, we illuminated it by three lighting spectra picked from LSPDD database. We plot the measured errors against the ground truth constituent images for 26 rendered scenes.

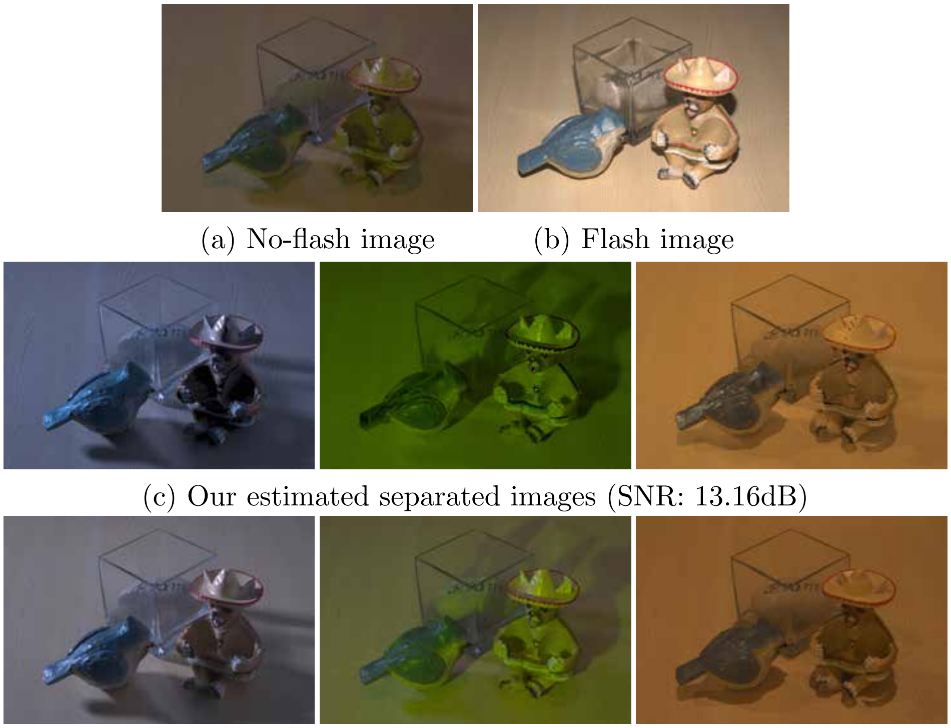
applying NNMF to the no-flash image leads to the loss of the colors. Hsu et al. [77] use the no-flash photograph to estimate the relative contribution of the light sources by introducing restrictive assumptions on the scene as well as the colors of the illuminants; while we manually selected the light colors to guide the reconstruction of this technique, there are numerous visual artifacts due to the use of strong scene priors. In contrast, our technique produces results that closely resemble the actual captured photographs, indicating its robustness and effectiveness.



(a) Input images(b) Matrix factorization(c) Hsu et al. [77](d) Our results(e) Ground truth
SNR 16.96 dB SNR 10.13 dB SNR 20.43 dB

Figure 4.8: We separate a no-flash image (a) into two components and compare with matrix-factorization (b) and Hsu et al. [77] (c). Compared to the ground truth images, we can see that matrix factorization produces noisy colors (see the painting on the left), while Hsu et al. [77] produce an incorrect estimate of light color and shading. Our result (d) closely mimics the actual captured results.

Scenes with three lights. The proposed technique is, to our knowledge, the first to demonstrate three light source separation. In Figure 4.9, we compare our technique to the ground truth on scenes with three lights. The scene is illuminated under warm indoor lighting, a green fluorescent lamp and cool skylight. Our lighting separation scheme produces visually pleasing results with shadows and shadings that are consistent with those observed in the ground truth. Figure 4.10 showcases separation on two additional scenes. For the scene in the top row, our technique for estimating lighting coefficients fails due to lack of shadows; to obtain the separation, we had to manually pick the corners of $\mathcal{G}$ to estimate the illumination coefficients.



(a) No-flash image (b) Flash image

(c) Our estimated separated images (SNR: 13.16dB)

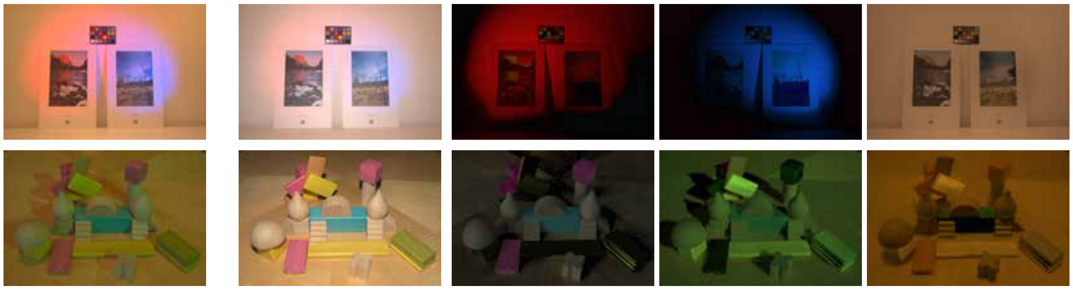
(d) Captured photographs

Figure 4.9: We evaluate our technique on scenes with mixtures of three lights and compare with the ground truth image. Our technique is able to capture both the color and the shading for each of these sources and produce results similar to the ground truth.

4.5.2 Evaluation of reflectance estimation

We report the performance of our reflectance estimation on both synthetic datasets and real-world scenes.

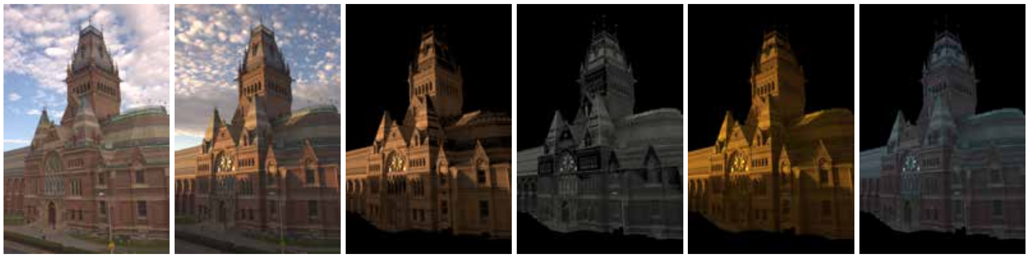
Synthetic experiments. We use the MIT database [17] together with the rendered scenes with ground truth to evaluate the performance of our technique. The examples are shown in Figure 4.5, 4.13, respectively. For the each image in MIT database, we generate the no-flash images by using the



(a) No-flash images (b) Flash images (c) Estimated separated images

Figure 4.10: We evaluate our technique on scenes with three lights. (top row) We capture an image under warm indoor LED lights and two LED lights with red and blue filter, respectively. Our technique is able to estimate separated results that capture this complex light transport. (bottom row) We image a scene under warm indoor lighting, a green fluorescent lamp and cool skylight. Our separation results capture both the color and the shading for each of these sources.

surface normals of each object and modulating with two point light source, each with lighting spectrum in LSPDD database [2]. The corresponding ground truth shading are generated by using the dot product of surface normal and lighting direction of each light source. To generate the pure flash image, we use the point light source in the frontal view, i.e. the same as the view direction. We use the no-flash/flash pair to show the performance of the proposed technique, and also compare with the state-of-the-art techniques. In particular, we evaluate against the baseline method by using color retinex, single-image techniques of Shen et al. [135], Bell et al. [20], Li et al. [94], Zhao et al. [159], Barron and Malik [17], as well as the multi-images techniques of Weiss [151], Weiss + Retinex [66] and Hauagge et al. [68]. We use the no-flash image as the input to test the performance of single-image



Frame 40 (pure flash) Frame 16 Prinet et al. [120] Our results

Figure 4.11: Sun- and Sky-light separation. We use photo on a cloudy day as the pure flash image. Note that the Sun being a directional light source casts sharp shadows onto the scene, while the Sky being an area light does not induce shadows. As can be seen from the separated images, our algorithm is able to produce good results with convincing color and shading attributes for both sources in the scene.

techniques, and the no-flash/flash image pair for the multi-image techniques. We characterize the performance of both reflectance and shading estimation by testing on the synthetic data with varying objects in the database. We use the relative GMSE to quantify the accuracy of the estimate, where it is defined as

$$\mathrm{GMSE}(x, \widehat{x}) = \|x - \widehat{\alpha}\widehat{x}\|_2^2,$$

with $\widehat{\alpha} = \arg\min_\alpha \|x - \alpha\widehat{x}\|_2^2$. That is, GMSE evaluates the results by compensating the brightness difference in the separation. As can be seen from Figure 4.12, we are able to produce results that are, in terms of GMSE error, an order of magnitude better than the state-of-the-art methods.

Though our model performs well on the MIT database [17] for single object, scenes with multiple objects present a more challenging scenario. To illustrate this, we compare our method with the concurrent work on the scenes involving with more complicate objects by using MITSUBA rendering engine [85]. A few examples are shown in Figure 4.14. The rendering images place a layer of difficulty to handle the interaction of illumination with the objects, including inter-reflection, cast shadows. For each scene, we select two lighting spectrum from LSPDD [2] database to render the image. In Figures 4.13 and 4.14, we showcase the performance against the state-of-the-art techniques. As can be seen from Figure 4.13, most of the techniques assume a single light color or one ambient white illuminant, which leads to color variations in the shading to be incorporated into the reflectance estimates. Unlike these approaches, our technique handles spatially-varying color illumination, and correctly separates the detail in the shading and reflectance layers.

Real scenes. Real images present a layer of difficulty well beyond simulations and introduce inter-reflections, sub-surface scattering, cast shadows, and imprecise light source localization. We test the performance of our reflectance and shading estimation algorithm on a wide range of real world scenes. Figure 4.15 visually illustrates the performance of our technique for the reflectance and shading on two real-world scenes and compares with state-of-the-art techniques. The performance of these techniques on the real-world scenes closely parallels the results we observed in the synthetic dataset. As can be seen in Figure 13, our technique significantly outperforms the state-of-the-art methods in the estimates for the reflectance and shading. While

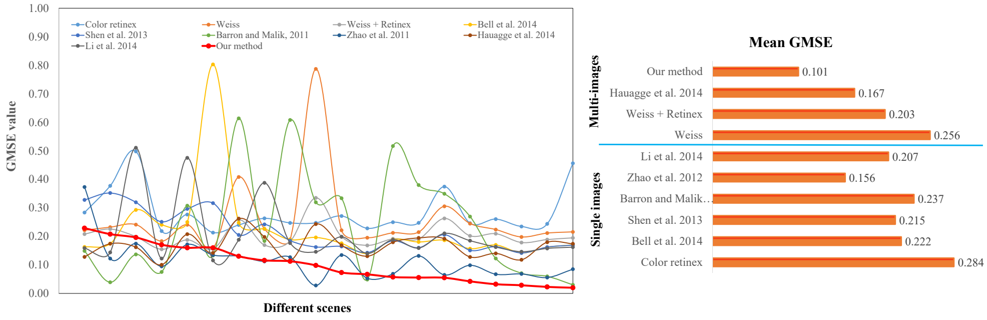


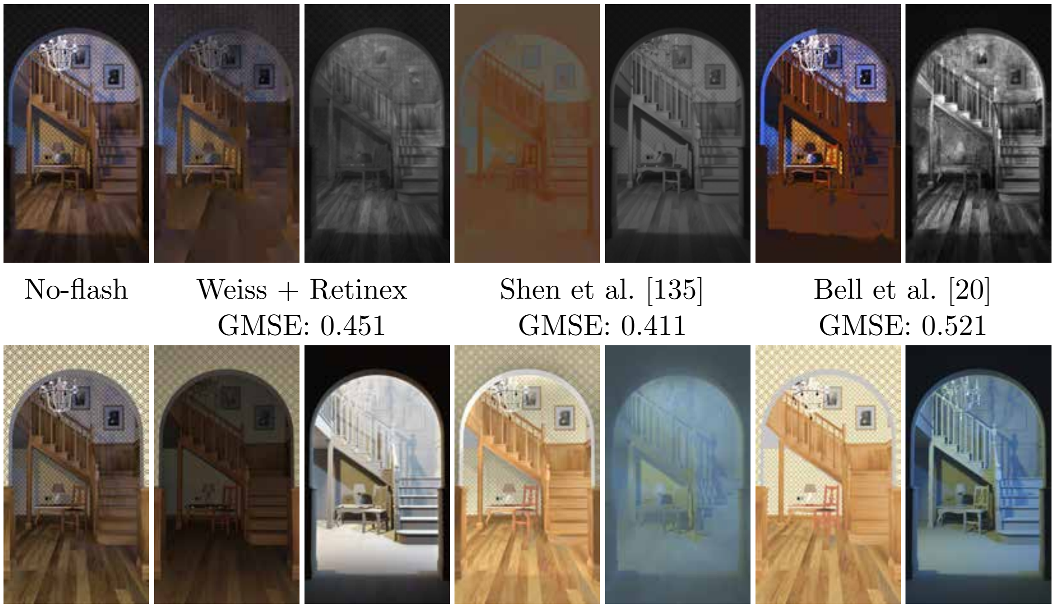
Figure 4.12: We evaluate the performance of different intrinsic image algorithms on synthetic objects from the MIT database [66] when imaged under two illuminants. For each object, we compute GMSE compared to the ground truth. The plot (right) shows the mean error over all the test objects.

our technique requires one additional image, the only intrinsic image techniques, that we are aware of, that can handle these situations either require a larger set of images or significant user interaction. In contrast, our results are completely automatic.

4.5.3 Applications

Source separation of the form proposed is invaluable in many applications. We consider six distinct applications: white balancing under mixed illumination, post-capture editing of illuminant spectrum and brightness, sun/sky-light separation, manipulation of camera spectral response, two-shot photometric stereo and texture editing.

White balancing under mixed illumination. One of the applications enabled by our technique is white balancing under mixed illumination. The



No-flash

Weiss + Retinex
GMSE: 0.451

Shen et al. [135]
GMSE: 0.411

Bell et al. [20]
GMSE: 0.521

Flash

Hauagge et al. [68]
GMSE: 0.326

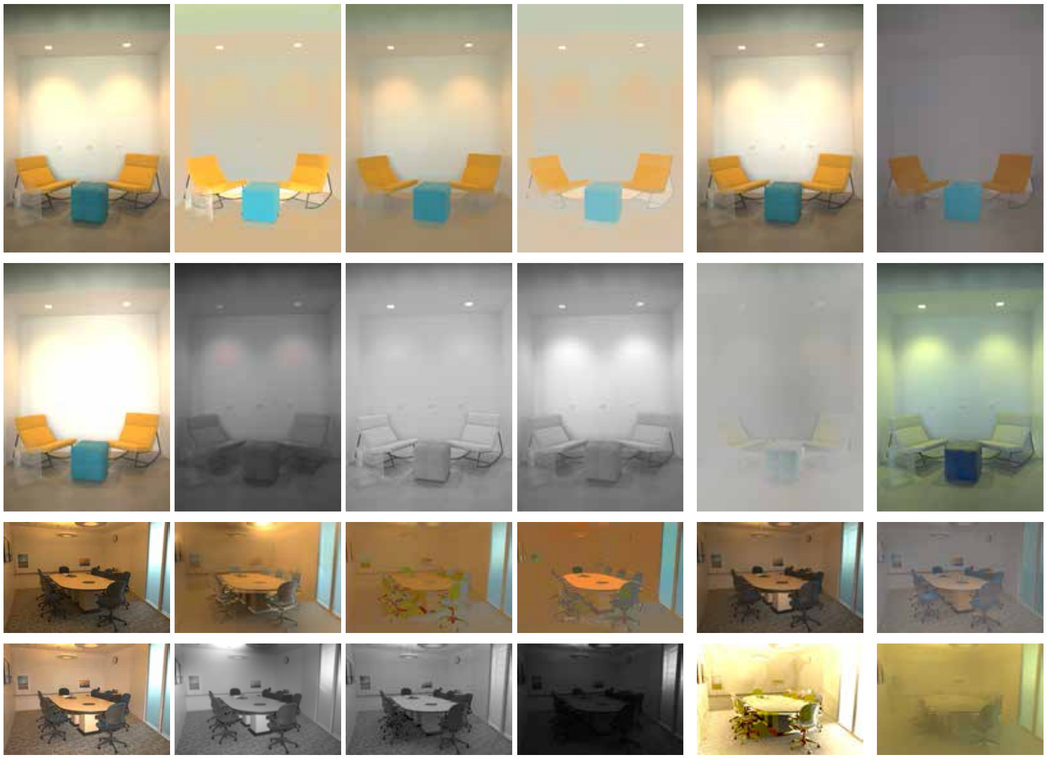
Our method
GMSE: 0.185

Ground truth

Figure 4.13: We evaluate the performance of our method against the state-of-the-art methods on a rendered scene with mixture of two light sources. In parallel to the results observed in 4.14, our method achieves the best performance in term of GMSE. In addtion, as can be seen here, our technique is able to attribute the illuminant brightness variation and induced shadows into the shading layer while cluster the texture on the floor to the reflectance layer.

	Methods	Kitchen	Living-room	Living-room2	Dinning-room
Single image	Color retinex	0.732	0.571	0.679	0.511
	Bell et al. 2014	0.591	0.428	0.547	0.401
	Shen et al. 2013	0.486	0.392	0.478	0.384
	Barron and Malik 2011	0.691	0.530	0.602	0.534
	Zhao et al. 2012	0.457	0.316	0.413	0.283
	Li et al. 2014	0.520	0.408	0.474	0.424
Multi-image	Weiss	0.621	0.501	0.598	0.448
	Weiss + Retinex	0.573	0.475	0.531	0.391
	Hauagge et al. 2014	0.412	0.219	0.399	0.219
	Our method	0.258	0.150	0.216	0.199

Figure 4.14: We evaluate the performance of different intrinsic image algorithms on synthetic objects from the rendering images by MITSUBA [85] under two illuminants. For each scene, we compute GMSE compared to the ground truth and marked the best performance in red. As can be seen here, the proposed technique is able to achieve best performance on all the test scenes.



No-flash/ flash images

Weiss + Retinex [66]

Shen et al. [135]

Bell et al. [20]

Hauagge et al. [68]

Our results

Figure 4.15: We compare our estimated reflectance (row 1,3) and shading (row 2,4) with state-of-the-art techniques. As can be seen here, our results are significantly better than other technique. We handle spatially-varying lighting, which is often baked into the reflectance for other methods (see the light highlights on the wall, and shadows on the floor). We also accurately separate texture from the shading (see the texture on the carpet and the text on the table).

vast majority of white-balance algorithms assume that the scene is lit by a single dominant light source. In contrast, we are able to estimate and remove the effect of spatially-varying lighting by using the Hull constraint. There are two approaches to achieve this.

- **Approach I.** We can simply adjust the illumination coefficients in each separated images. In particular, we substitute the estimated coefficients $\widehat{\mathbf{b}}_j$ in (4.18) with the coefficients corresponding to the neutral light spectra. However, this approach requires us to estimate the light source coefficients and their relative shading, which can only be performed for 3 or fewer light sources.

- **Approach II.** We provide an alternative solution that provides the ability to handle any number of light sources in the scene, albeit under some assumptions on their colors. Specifically, we assume that $\| \sum_{i=1}^{N} \eta_i \mathbf{b}_i \|^2 \approx \| \sum_{i=1}^{N} \eta_i \|^2$. That is, $\sum_{i \neq j} \eta_i \eta_j \mathbf{b}_i \mathbf{b}_j \approx \sum_{i \neq j} \eta_i \eta_j$, or equivalently, $\mathbf{b}_i \mathbf{b}_j \approx 1$. In essence, we have constrained the lighting spectra close to each other in the low-dimensional model.

  Now, recall that the no-flash intensity is

$$I_{\mathrm{nf}}^k(\mathbf{p}) = \|\mathbf{a}_\mathbf{p}\| \left(\frac{\alpha_\mathbf{p}^T}{\|\alpha_\mathbf{p}\|} \right) E^k \ \sum_{i=1}^{N} \eta_i(\mathbf{p}) \mathbf{b}_i, \tag{4.15}$$

the white balancing results at pixel $\mathbf{p}$ can be expressed as

$$I_{\mathrm{wb}}^{k}(\mathbf{p}) = \|\mathbf{a}_{\mathbf{p}}\| \left(\frac{\alpha_{\mathbf{p}}^{T}}{\|\alpha_{\mathbf{p}}\|} \right) E^{k}\ \mathbf{b}_{\mathrm{wb}} \sum_{i=1}^{N} \eta_i(\mathbf{p}), \tag{4.16}$$

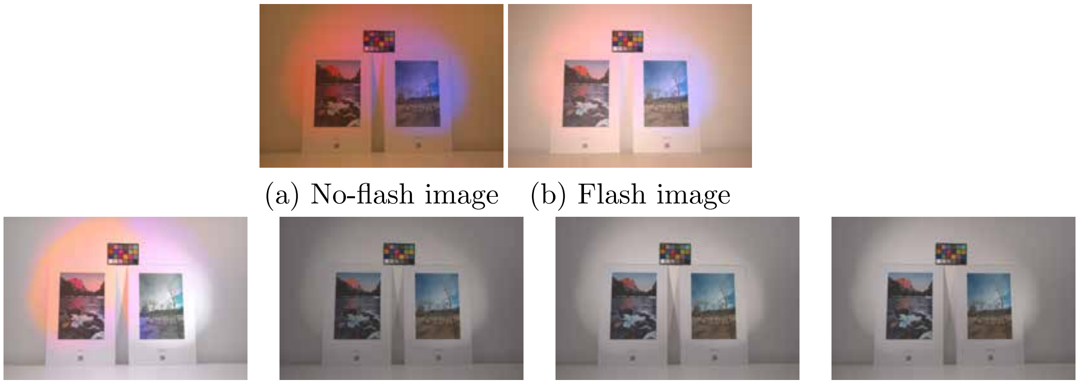
where $\mathbf{b}_{\mathrm{wb}}$ is the neutral light coefficients.

Given $\| \sum_{i=1}^{N} \eta_i \mathbf{b}_i \|^2 \approx \| \sum_{i=1}^{N} \eta_i \|^2$, we can substitute $\|\beta(\mathbf{p})\|$ to express (4.16) as

$$I_{\mathrm{wb}}^{k}(\mathbf{p}) = \left(\frac{\alpha_{\mathbf{p}}^{T}}{\|\alpha_{\mathbf{p}}\|} \right) E^{k} \|\beta(\mathbf{p})\| \mathbf{b}_{\mathrm{wb}}$$

Our results are shown in Figures 4.16. We compare our results with those from two algorithms that are designed to handle spatially-varying mixed illumination — Hsu et al. [77] and Hui et al. [80]. Hsu et al. require that the color of the illuminants and assume that only two light sources present in the scene. While we manually specified this as input to their technique, their result is not able to deal with extreme illumination (see Fig. 4.16). Similar to our work, Hui et al. use a flash camera and can generalize to an arbitrary number of scene illuminants. While Hui et al. [80] produces the results of similar quality, their underlying image formation model is not physically accurate; in particular, their method completely ignores the image formation model in 4.1 and instead uses an empirical model that does not account for camera spectral response.

Sunlight and skylight separation. An interesting application of two-light source separation is in outdoor time lapse videos where it is often nec-

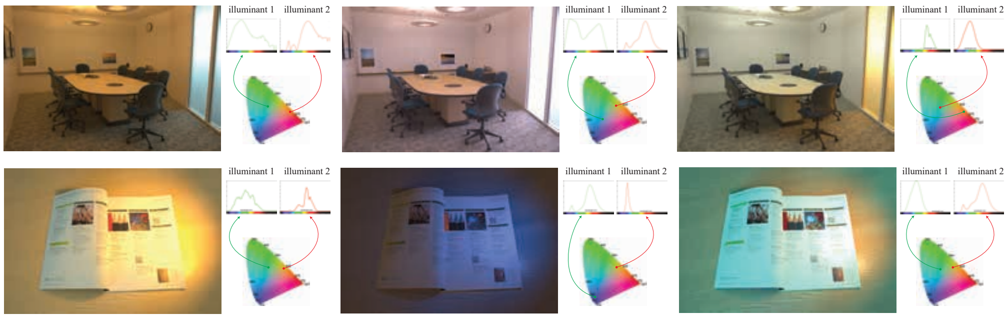


(a) No-flash image (b) Flash image

(c) Hsu et al. [77] (Mean error = 3.58°) (d) Hui et al. [80] (Mean error = 0.88°) (e) Our result (I) (Mean error = 0.85°) (f) Our result (II) (Mean error = 0.92°)

Figure 4.16: We evaluate our white balance method on a no-flash/flash pair (a/b) from Hui et al. [80]. We compare the white balance results with both Hsu et al. [77] and Hui et al. [80]. (c) Hsu et al. [77] require the light colors to be manually specified but fail on the extreme illumination in this scene. (d) Hui et al. use a flash image to improve results but rely on inaccurate physical model. (e) Our method (both approach I (e) and approach II (f)) produces the result which can achieve the same performance in terms of both visual quality and angular error. Note that approach (II) is able to produces visual appealing results as well as similar angular error measurements, making our method applicable to arbitrary number of light sources in the scene.

essary to separate direct sunlight from indirect skylight. Figure 4.11 showcases the performance of light separation technique on an outdoor scene. We identify a photograph with cloudy sky, where there is no direct sunlight and the entire scene is lit only by the skylight, as a pure flash photograph. Since our technique does not make any assumptions about the nature of the flash illumination, we use skylight in place of the flash light. Also note that skylight changes its color and intensity significantly during the course of the day. Given this pure flash photograph, our separation scheme is able



(a) No-flash image (b) Our light editing results

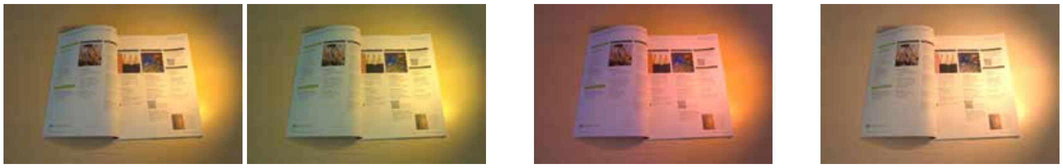
Figure 4.17: We separate no-flash images (a) into individual light components, and recolor them to create photo-realistic results with novel lighting conditions (b). We show the novel spectral distribution as well as the CIE plots for the light sources. Note how our method changes the color and brightness of each light while realistically retaining all shading effects.

to produce the results closely resemble to the manner of the sky and the sun illumination. We compare our method with the video-based work of Prinet et al. [120] on the time-lapse video sequence. While the method by Prinet et al. does not require the pure flash image, it assumes that the colors of the illuminants will not change which leads to artifacts in the separated images.

Post-capture manipulation of light color and brightness. Given the separated results, we can adjust the brightness as well as the spectrum of a particular light. Specifically, we can produce the photograph

$$\widetilde{I} = \sum_j \|\boldsymbol{\beta}(\mathbf{p})\|_2 \boldsymbol{\alpha}_{\mathbf{p}}^T E^k \widehat{z}_j(\mathbf{p}) \mu_j \widetilde{\mathbf{b}}_j, \tag{4.17}$$

where $\widetilde{\mathbf{b}}_j$ denotes the adjusted illumination coefficients and μ_j denotes the changes in the brightness. Figure 4.17 shows an example of editing the light



Nikon D700 Canon 5D Point Grey Grasshopper2 Nokia N900

Figure 4.18: Results on camera response editing. We show estimated rendering results for different camera models.

color and brightness for the captured no-flash images. We experiment by adjusting the parameters μ_j and $\widetilde{\mathbf{b}}_j$ in (4.17). The rendered photographs are both visually pleasing and photo-realistic in their preservation of shading and shadows.

Manipulation of camera spectral response. Another unique capability of the proposed method is its ability to edit camera spectral response. We can achieve this as follows. Given the estimate of the separated image as

$$\widehat{I}^k_{sep,j}(\mathbf{p}) = \|\beta\| \alpha_{\mathbf{p}}^\top E^k \widehat{z}_j(\mathbf{p}) \widehat{\mathbf{b}}_j. \tag{4.18}$$

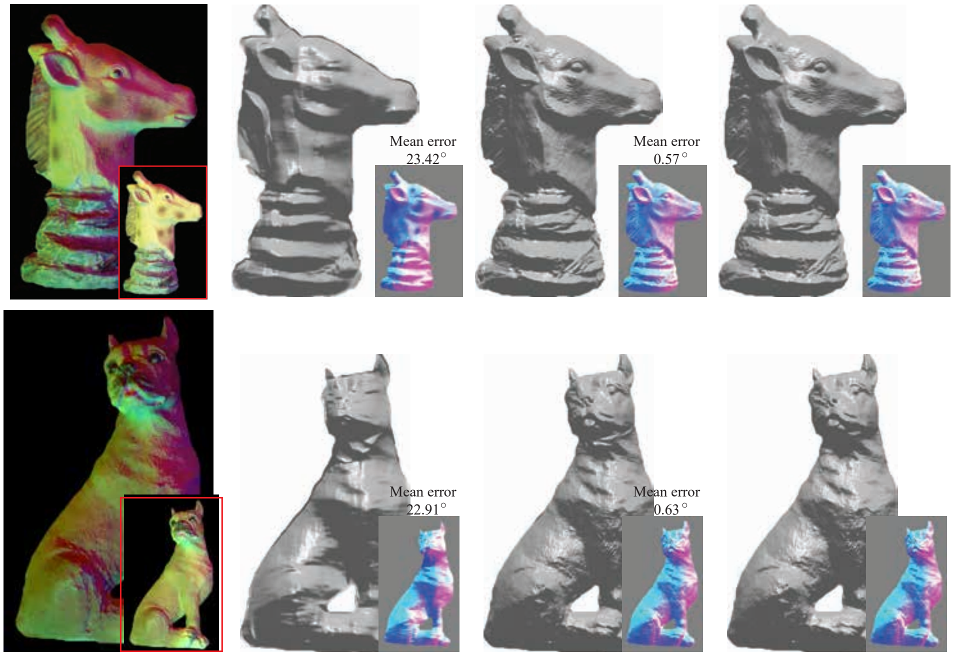
and

$$E^k(i,j) = \int_\lambda \widetilde{\rho}_i(\lambda) S^k(\lambda) \widetilde{\ell}_j(\lambda) d\lambda,$$

we are able to change the camera spectral response $S^k(\lambda)$ with a novel spectral distribution function $\widehat{S}^k(\lambda)$. Specifically, we change the captured no-flash image with novel camera response function and show the results in Figure 4.18.

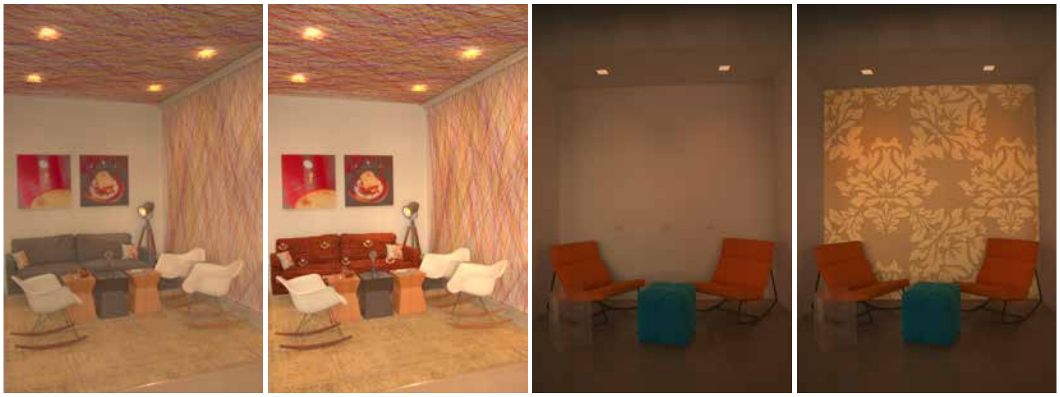
Flash/no-flash photometric stereo. Photometric stereo [140,152] methods aim to surface shape (usually normals) of an object from images obtained from a static camera under varying lighting. For Lambertian objects, this requires a minimum of three images. Recently, techniques have been proposed to do this from a single shot where the object is lit by three monochromatic red, green, and blue, directional light sources [28, 32]. However this estimation is still ill-posed and requires additional priors. We propose augmenting this setup by capturing an additional image lit by a flash collocated with the camera. We use our proposed technique for source separation to create three images (plus the pure flash image), at which point we can use standard calibrated Lambertian photometric stereo to estimate surface normals. As shown in Figure 4.19 this leads to results that are orders of magnitude more accurate than the state-of-the-art technique [32]. More comparisons can be seen in the supplementary material.

Texture editing. Given the estimated reflectance and shading, we are able to edit on top of the estimates. In particular, we can operate on the reflectance layer to enable the texture editing of the scene. Figures 5.1 and 4.20 showcase the performance of texture editing on two real scenes. Note that the texture editing always placed a challenge for the state-of-the-art methods due to the underlying ambiguities in the intrinsic image decomposition. As can be seen here, our technique is able to produce visually appealing results and align well with the ambient illumination.



(a) RGB image pair (b) Results of [32] (c) Our result (d) Ground truth normals

Figure 4.19: Results on two-shot captured photometric stereo of real objects. We show estimated normal map for our technique as well as that of single-shot method of Chakrabarti et al. [32]. We include the mean of the angular errors for the estimated surface normals.



Original image Edited result Original image Edited result

Figure 4.20: Results on texture editing of a real scene. We show edited results for our technique on the real scene. As can be seen here, the edit texture on the sofa aligns well the ambient illumination, which demonstrate the robustness of our technique on intrinsic image decomposition.

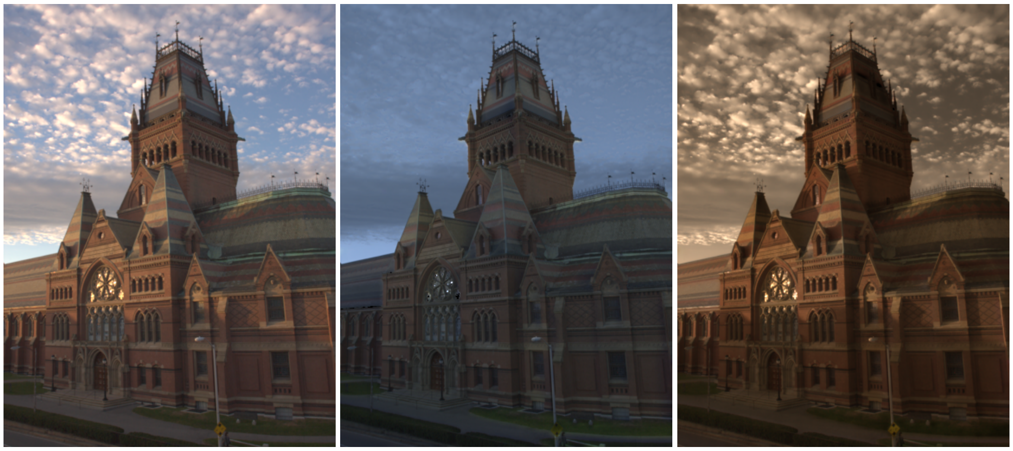
Chapter 5

Single photograph illumination analysis

Natural environments are often lit by multiple light sources with different illuminant spectra. Depending on scene geometry and material properties, each of these lights causes different light transport effects like color casts, shading, shadows, specularities, etc. An image of the scene combines the effects from the different lights present, and is a superposition of the images that would have been captured under each individual light. We seek to invert this superposition, i.e., separate a single image observed under two light sources, with different spectra, into two images, each corresponding to the appearance of the scene under one light source alone. Such a decomposition can give users the ability to edit and relight photographs, as well as provide information useful for photometric analysis.

However, the appearance of a surface depends not only on the properties of the light sources, but also on its spatially-varying geometry and material properties. When all of these quantities are unknown, disentangling them is a significantly ill-posed problem. Thus, past efforts to achieve such separation have relied heavily on extensive manual annotation [24–26] or access to calibrated scene and lighting information [41, 42]. More recently, Hui et al. [81] demonstrate that the lighting separation problem can be reliably solved if one additionally knows the reflectance chromaticity of all surface points — which they recover by capturing a second image of the same scene under flash lighting. Given that the flash image is used in their processing pipeline only for estimating the reflectance chromaticity, could we computationally estimate the reflectance chromaticity from a *single image*, thereby avoiding the need to capture a flash photograph all together? This would greatly enhance the applicability of the method especially for scenarios where it is challenging to sufficiently illuminate every pixel with the flash (e.g., when the flash is not strong enough, the scene is large, or the ambient light sources are too strong).

Our work is also motivated by the success of deep convolutional neural networks for solving closely related problems like intrinsic decompositions [95, 160], and reflectance estimation [96, 124, 148]; hence, we propose training a deep convolution neural network to perform this separation. However, we find that standard architectures, trained only with the respect to the quality of the final separated images, are unable to learn to effectively



(a) Input image (b) Output separated images

Figure 5.1: Our method separates a single image (a) captured under two illuminants with different spectra (sun and sky illumination here) into two images corresponding to the appearance of the scene under the individual lights. Note that we are able to accurately preserve the shading and shadows for each light.

perform the separation. Therefore, we guide the design of our network using a physics-based analysis of the task [81] to match the expected sequence of inference steps and intermediate outputs — reflectance chromaticities, shading chromaticities, separated shading maps, and final separated images. In addition to ensuring that our architecture has the ability to express these required computations, this decomposition also allows us to provide supervision to intermediate layers in our network, which proves crucial to successful training.

We train our network on two existing datasets: the synthetic database of Li et al. [95], and the set of real flash/no-flash pairs collected by Aksoy et al. [8] — using a variant of Hui et al.'s algorithm [81] to compute ground-truth values. Once trained, we find that our approach is able to successfully solve this ill-posed problem and produce high-quality lighting decompositions that, as can be seen in Figure 5.1, capture complex shading and shadows. In fact, our network is able to match, and in specific instances outperform, the quality of results from Hui et al.'s two-image method [81], despite needing only a single image as input.

5.1 Related Work

Estimating illumination and scene geometry from a single image is a highly ill-posed problem. Previous work has focused on specific subsets of this problem; we discuss previous works on illumination analysis as well as prior attempts of intrinsic image decomposition that aim to jointly estimate the illumination and surface reflectance.

Illumination estimation. Estimating the ambient illumination from a single photograph has been a long-standing goal in computer vision and computer graphics. The majority of past techniques have been extensive studied in literature of color constancy [61] — the problem of removing the color casts of ambient illumination. One popular solution is to model the scene with single dominant light source [52, 55, 60]. To deal with the mixture of lightings in the scene, previous works [45, 62, 126] typically characterize each local region with different but single light source. However, these approaches cannot generalize well to scenes where multiple light sources mix gracefully. To address this, Boyadzhiev et al. [27] utilize user scribbles to indicate color attributes of the scene such as white surfaces and constant lighting regions. Hsu et al. [77] propose a method to address mixtures of two light sources in the scene; however, they require the precise knowledge of the color of each illuminant. Prinet et al. [120] resolve the color chromaticity of two light sources by utilizing the consistency of the reflectance of the scene in a sequence of images. Sunkavalli et al. [144] demonstrate this (and image separation) for time-lapse sequences of outdoor scenes.

In parallel, many techniques have been developed to explicitly model the illumination of the scene, rather than removing the color of the illuminants. Lalonde et al. [92] propose the parametric model to characterize the sky and sun for the outdoor photographs. Hold-Geoffroy et al. [75] extend the idea to model the outdoor illumination by incorporating a deep neutral network. Gardner et al. [59] utilize a data-driven approach to represent the indoor illumination from a single LDR photograph. In contrast, our method does not model the illumination in certain form, but directly regresses the single-illuminant images.

Intrinsic image decomposition. Intrinsic image decomposition methods seek to separate a single image into a product of reflectance and illumination layers. This problem is commonly solved by assuming that reflectance of the scene is piece-wise linear while the illumination shading varies smoothly [18].

Several approaches developed along this line by further imposing priors on non-local reflectance [20, 135, 159], on the consistency of reflectance for the images sequence taken under static camera [66, 68, 90]. A common assumption in intrinsic image methods is that the scene is lit by a single dominant illuminant. This does not generalize to real world scenes with visually complex object and illuminated with mixture of multiple light sources. Recently, the vast majority of techniques [95, 97] are devoted to deep neutral networks with large amount of data to improve the conditioning of the problem. While effective, these techniques also focus on the scene illuminated with single light source. Barron et al. [15, 17] resolve this by incorporating the global lighting model to characterize spatially-varying illumination. While this lighting model works well for single object, it is unable to capture high-frequency spatial information, like shadows that are often present in real scenes. In comparison, our technique is well-suited for the scene with mixture of multiple light sources and able to work well for the scenes with complex geometry. In addition, as opposed to predicting the reflectance of the scene, our method only requires to predict its *chromaticity*, which is an easier problem to solve.

5.2 Problem Statement

Our objective is to take as input, a single photograph of a scene lit by a mixture of two illuminants, and estimate the images lit by each single light source. This is a severely ill-posed problem and we propose solving it using

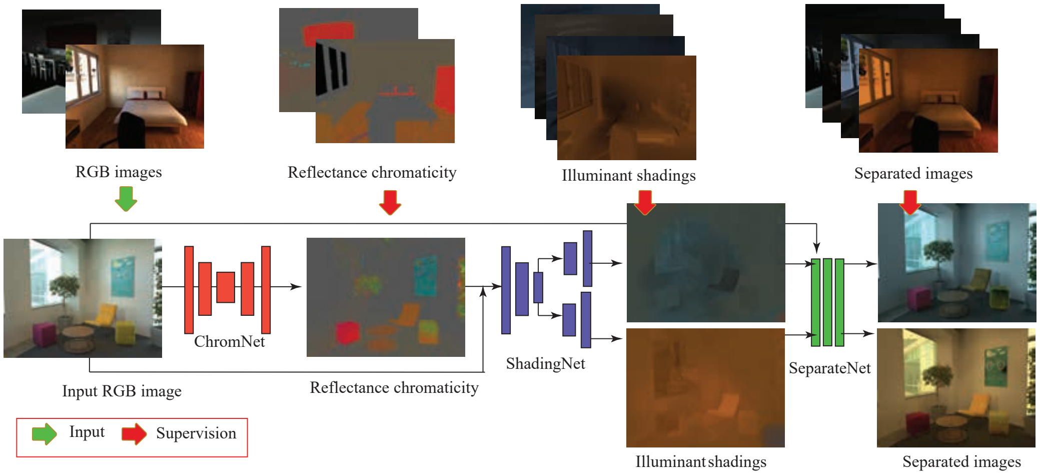


Figure 5.2: Given a single image under the mixture of lighting, our method automatically produces the images lit by each of these illuminants. We train a cascade of three sub-networks with three specific tasks. First, we estimate the reflectance color chromaticity of the scene via ChromNet. Given this estimation, we concatenate it with the input RGB image and feed them into ShadingNet to predict the illuminant shadings. We append these to the input image and pass it to SeparateNet to produce the output. During training, we supervise the reflectance chromaticity, illuminant shadings and the separated images.

deep neural networks. In this section, we set up the image formation model and describe the physical priors we impose to supervise the intermediate results produced by the network.

5.2.1 Problem setup and image formation

We adopt the image formation model from Hui et al. [81] by assuming that the scene is Lambertian and is imaged by a three-channel color camera. However, instead of modeling infinite-dimensional spectra using subspaces, we assume that the camera color response is narrow-band, allowing us to characterize both the light source and albedo in RGB. That is, the intensity observed at a pixel $\mathbf{p}$ in a single photograph I is given by:

$$I^c(\mathbf{p}) = R^c(\mathbf{p}) \sum_{i=1}^{N} \lambda_i(\mathbf{p})\ \ell_i^c, \quad \text{for } c \in \{r, g, b\}, \tag{5.1}$$

where $R(\mathbf{p}) = [R^r(\mathbf{p}), R^g(\mathbf{p}), R^b(\mathbf{p})]$ is the three-color albedo. In our work, we focus on the scenes that are lit by $N = 2$ light sources and we denote the light chromaticities as $\{\boldsymbol{\ell}_1, \boldsymbol{\ell}_2\}$. Note that $\boldsymbol{\ell}_i = [\ell_i^r, \ell_i^g, \ell_i^b] \in \mathbb{R}^3$ with $\sum_x \ell_i^x = 1$. Similar to Hui et al. [81], we assume that the light source chromaticities are unique, i.e., $\boldsymbol{\ell}_1 \neq \boldsymbol{\ell}_2$. The term $\lambda_i(\mathbf{p})$ is the shading observed at pixel $\mathbf{p}$ due to the i-th light source multiplied by the light-source brightness. Given the fact that two sources with the same color are clustered together, the shading term $\lambda_i(\mathbf{p})$ has a complex dependence on the lighting geometry and does not have a simple analytical form. Our goal is to compute the separated images corresponding to the each light source k as:

$$\widehat{I}^c_{\text{sep},k}(\mathbf{p}) = R^c(\mathbf{p})\ \lambda_k(\mathbf{p})\ \ell_k^c. \tag{5.2}$$

To solve this, Hui et al. [81] capture additional image under flash illumination, that is used to compute the reflectance color chromaticity of the scene, from which it is able to isolate the reflectance from the illumination shading. Given the reflectance invariant space, they solve for the lighting colors of each light source as well as the per-pixel contribution of each illuminant. We provide a quick summary of the key steps of their computational pipeline, adapted to the RGB color model.

Step 1 — Flash to reflectance chromaticity. The pure flash photograph enables us to estimate the reflectance chromaticity $\alpha^c(\mathbf{p})$ defined as

$$\alpha^c(\mathbf{p}) = \frac{R^c(\mathbf{p})}{\sum_x R^x(\mathbf{p})}. \tag{5.3}$$

Step 2 — Estimate shading chromaticity. Using the reflectance chromaticity $\alpha^c(\mathbf{p})$, we next derive the shading chromaticity $\gamma^c(\mathbf{p})$ defined as

$$\gamma^c(\mathbf{p}) = \frac{\sum_i \lambda_i(\mathbf{p})\ \ell_i^c}{\sum_j \lambda_j(\mathbf{p})} = \sum_{i=1}^{N} z_i(\mathbf{p})\ \ell_i^c, \tag{5.4}$$

where we denote

$$z_i(\mathbf{p}) = \frac{\lambda_i(\mathbf{p})}{\sum_{j=1}^{N} \lambda_j(\mathbf{p})} \tag{5.5}$$

as the relative shading term. As indicated by Hui et al. [81], γ is key for estimating the relative shading from the illumination shadings, from which we are able to separate the images with respect to the illuminant colors.

Step 3 — Estimate relative shading. From the shading chromaticity, the illuminant shadings S_i^c for each light source is

$$S_i^c(\mathbf{p}) = z_i(\mathbf{p})\ell_i^c. \tag{5.6}$$

We can now get the separated images using the following expression:

$$\widehat{I}^{c}_{\text{sep},k}(\mathbf{p}) = I^{c}(\mathbf{p})\frac{S^{c}_{k}(\mathbf{p})}{\sum_{i=1}^{N} S^{c}_{i}(\mathbf{p})}. \tag{5.7}$$

In this chapter, we design our network by mimicking the steps in the derivation above, but each processing element is replaced with deep networks as shown in Figure 5.2. In particular, we utilize three sub-networks — ChromNet, ShadingNet and SeparateNet — to estimate the reflectance chromaticity, illuminant shadings and separated images, respectively. ChromNet predicts the values of reflectance chromaticity α, defined in (5.3), with its input being the RGB image that we seek to separate. ShadingNet takes in as the output of ChromeNet concatenated with the input RGB image to regress the illuminant shadings in (5.6). Finally, SeparateNet gathers the estimated illuminant shadings as well as the input RGB image to estimate the separated images.

5.2.2 Generating the training dataset

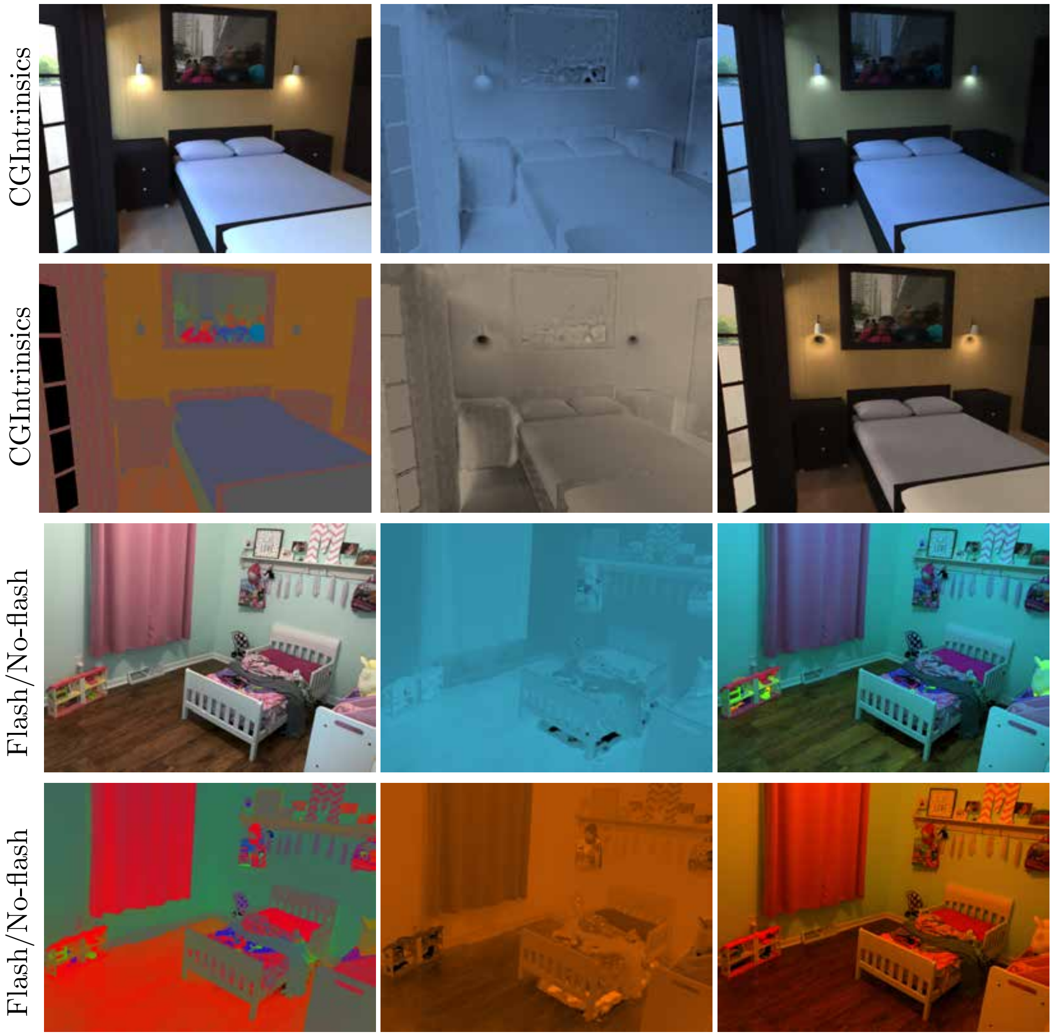
We utilize the databases of CGIntrinsics [95] and Flash/No-Flash [8] to produce (approximate) ground truth reflectance chromaticity, illuminant shadings and separated images. Figure 5.3 shows an example of the training data from each dataset.

The CGIintrinsics dataset consists of 20160 rendered scenes from SUNCG [141] and provides the ground truth reflectance, from which we compute the reflectance chromaticity. We then estimate the shadings chromaticity by using (5.4).

The Flash/No-flash dataset consists of 2775 image pairs. We estimate the reflectance chromaticity as the color chromaticity of the pure flash image, which is the difference between the flash and the no-flash photograph. We anecdotally observed that the majority of the scenes in this dataset are only illuminated by a single light source — which, as such, makes it uninteresting for our application. To resolve this, we add the flash image back to no-flash image and create photographs illuminated by two light sources. By changing the color of the flash photograph, we can enhance the amount of training data; this allows us to generate 29060 input-output pairs, where the input is a photo, and the output is the reflectance chromaticity, a pair of its corresponding illuminant shadings as well as the separated images.

5.3 Learning Illuminant Separation

Now that we have the training data for the intermediate results, i.e. reflectance chromaticity, relative shadings and separated images, we detail our approach for learning the relationship between a single photo and its constituent images lit by each illuminant.



(a) Input (Top) / Chromaticity (Bottom)

(b) Illuminant shadings

(c) Separated images

Figure 5.3: We showcase each sample of train pairs from CGIntrinsics (top) and Flash/No-flash database (bottom). Given the input photograph (a), we use reflectance chromaticity together with illuminant shadings (b) and separated images (c) to supervise the output of the network.

5.3.1 Network architecture

As shown in Figure 5.2, we use a deep neural network to match the computation of the separation algorithm in Section 5.2. Specifically, our network consists of three sub-networks that produce the reflectance chromaticity, illuminant shadings, and the separated images respectively.

ChromNet. We design the first sub-network to explicitly estimate the reflectance chromaticity (5.3) from the input color image. This essentially requires the network to solve the ill-posed problem of estimating and removing the illumination color cast given only a single photograph. We adopt an architecture similar to that of Johnson et al. [86] to map the input image to a three channel reflectance chromaticity map.[1]

ShadingNet. The second sub-network in our framework takes reflectance chromaticity estimates as inputs, and solves for the two illuminant shadings in (5.6). From Section 5.2, we expect the first part of this computation to involve deriving γ from the chromaticities and original input, on a purely per-pixel basis as per (5.4). However, we found computing the γ values explicitly to lead to instability in training, likely since this involves a division. Instead, we produce a general feature map intended to encode the γ information (note that we do not require it to exactly correspond to γ values): we concatenate

[1] A detailed description of the construction of each subnetwork is provided in the supplemental material. We will also release our code base, training data and trained models upon acceptance.

the input image with the estimated chromaticities, and include two 1 × 1 convolution layers to produce a 16-channel feature map.

Given this feature map, our second sub-network produces the two separated illuminant shading maps. Since this requires global reasoning, we use an architecture similar to the pixel-to-pixel network of Isola et al. [84] to incorporate a large receptive field. However, since we need to produce two outputs shading maps from a single input feature map, we retain their architecture for the encoder that maps the feature map to a coarse resolution bottleneck, and include two copies of the decoder each of which maps this coarse output a different three-channel illuminant shading map. Both decoders in this architecture receive skip-connections from intermediate layers of the encoder.

SeparateNet. Given the illuminant shadings and previously estimated reflectance chromaticity, the last computation step is to produce the separated images. Here again, we use a series of pixel-wise layers to express the computation in (5.7). Our third sub-network concatenates the two predicted shading maps and the input RGB photograph into a nine-channel input, and uses three 1 × 1 convolution layers to produce a six-channel output corresponding to the two final separated RGB images.

Note that the output of our first sub-network—reflectance chromaticity—is sufficient to perform separation using the method of Hui et al. [81]. However, training this sub-network based directly on the quality of reflectance chromaticity estimates proves insufficient, because the final separated image quality can degrade differently with different kind of errors in chromaticity estimates. Thus, our goal is to instead train the reflectance chromaticity estimation sub-network to be optimal towards final separation quality. Unfortunately, the separation algorithm in [81] has non-differentiable processing steps, as well as other computation that produces unstable gradients. Hence, we use two additional sub-networks to approximate the processing in Hui et al.'s algorithm [81]. However, once trained, we find it is optimal to directly use the reflectance chromaticity estimates with the exact algorithm in [81], over the output of these sub-networks.

5.3.2 Loss functions

ChromNet loss. For the reflectance chromaticity estimation task, we use a scale-invariant loss. We also incorporate ℓ_1 loss in gradient domain, to enforce that the estimated reflectance chromaticity is piece-wise constant. In particular, we define our loss function as

$$\mathcal{L}_\alpha = \frac{1}{M}\sum_{i=1}^{M} \|\alpha_i^* - c_\alpha \alpha_i\|_1 + \sum_{t=1}^{L} \frac{1}{M_t}\sum_{i=1}^{M_t} \|\nabla\alpha_{t,i}^* - c_\alpha \nabla\alpha_{t,i}\|_1, \tag{5.8}$$

where α^* denotes the predicted chromaticity, α is the ground truth provided, and c_α is a term to compensate for the global scale difference, which can be estimated via least squares. We also use mask to disregard the loss at pixels where we do not have reliable ground truth (e.g. pixels that are close to black or pixels corresponding to the outdoor environment map in the SUNCG dataset). M indicates the total number of valid pixels in an image. Similar to the approach of Li et al. [95], we include a multi-scale matching term, where L is the total number of layers specified (3 in the chapter) and M_t denotes the corresponding number of pixels not masked as invalid pixels.

ShadingNet loss. We impose an ℓ_2 loss on both the absolute value and the gradients of the relative shadings. This encourages spatially smooth shading solutions (as is commonly done in prior intrinsic images work). However, the network outputs two potential relative shadings and swapping these two predictions should not induce any loss. To address this, we define our loss function as

$$\mathcal{L}_S = \min\{\mathcal{L}_{S_{11}} + \mathcal{L}_{S_{22}}, \mathcal{L}_{S_{12}} + \mathcal{L}_{S_{21}}\}$$

where $\mathcal{L}_{S_{ij}}$ denote the loss between the i-th output with j-th illuminant shadings defined in (5.6). Specifically, $\mathcal{L}_{S_{ij}}$ is defined as $\mathcal{L}_{S_{ij}} = \mathcal{L}_{\text{data}(i,j)} + \mathcal{L}_{\text{grad}(i,j)}$, where

$$\mathcal{L}_{\text{data}(i,j)} = \frac{1}{M}\sum_{u=1}^{M} \|S^*_{i,u} - c_S S_{j,u}\|_2, \tag{5.9}$$

$$\mathcal{L}_{\text{grad}(i,j)} = \sum_{t=1}^{L} \frac{1}{M_t} \sum_{u=1}^{M_t} \|\nabla S^*_{i,t,u} - c_S \nabla S_{j,t,u}\|_2, \tag{5.10}$$

Here, S^*_i denotes the i-th illuminant shading prediction while S_j is the ground truth, and c_S is the global scale to compensate for the illuminant brightness.

SeparateNet loss. Our loss for the two separated images is similar to our ShadingNet loss:

$$\mathcal{L}_I = \min\{\mathcal{L}_{I_{11}} + \mathcal{L}_{I_{22}}, \mathcal{L}_{I_{12}} + \mathcal{L}_{I_{21}}\},$$

where $\mathcal{L}_{I_{ij}}$ is the ℓ_1 loss. Specifically, $\mathcal{L}_{I_{ij}}$ is defined as

$$\mathcal{L}_{I_{ij}} = \frac{1}{M} \sum_{u=1}^{M} \|I^*_{i,u} - c_I I_{j,u}\|_1, \tag{5.11}$$

where, I^*_i denotes the i-th separated image predication while I_j is the ground truth for the j-th light source, and c_I is scale factor for the global intensity difference.

5.4 Evaluation

We now present an extensive quantitative and qualitative evaluation of our proposed method. Please refer to our supplementary material for more details and results.

5.4.1 Test dataset

Synthetic benchmark dataset. To quantitatively evaluate our method, we utilize the high quality synthetic dataset of [23]. This dataset has approximate 52 scenes, each rendered under several different single illuminants. We first white balance each image of the same scene, and then modulate the white-balanced images with pre-selected light colors; these represent the ground truth separated images. The input images are then created by adding pairs of these separated images, each corresponding to one of the lights in the scene. We produce 400 test samples in the dataset and use both of the ground truth of reflectance chromaticity and separated results to evaluate our method.

Real dataset. We also evaluate the performance of our proposed technique on real images captured for both indoor and outdoor scenes. Specifically, we utilize the dataset of the indoor scenes collected by Hui et al. [81] as well as time-lapse videos for outdoor scenes. Hui et al. [81] capture a pair of flash/no-flash for the same scene. We take the no-flash images in the dataset as the input to the network. For the time-lapse videos, each frame serves as a test input as shown in Figure 5.1 (a).

Training details. We resize our training images to 384 × 512. We use Adam optimizer [89] to train our network with $\beta_1 = 0.5$. The initial learning rate is set to be 5×10^{-4} for all sub-networks. We cut down the learning rate by 1/10 after 35 epochs. We then train for 5 epochs with the reduced learning rate. We ensure that all our networks have converged with this scheme.

Error metric. We characterize the performance of our approach on both reflectance chromaticity and the separated images. We adopt the ℓ_1 error to quantitative measure the performance of the reflectance chromaticity. To

Methods	Chromaticity	Separated Images
Proposed		
Chrom-Only	**0.0308**	0.0398
Final-Only	—	0.0351
Full-Direct	0.0537	0.0288
Full+ [81]	0.0537	**0.0207**
SingleNet	—	0.0679
Shen et al. [135]	0.0821	0.0791
Bell et al. [20]	0.0785	0.0763
Li et al. [95]	0.0833	0.0821
†Hsu et al. [77]	—	0.0678
†Hui et al. [81]	—	0.0101

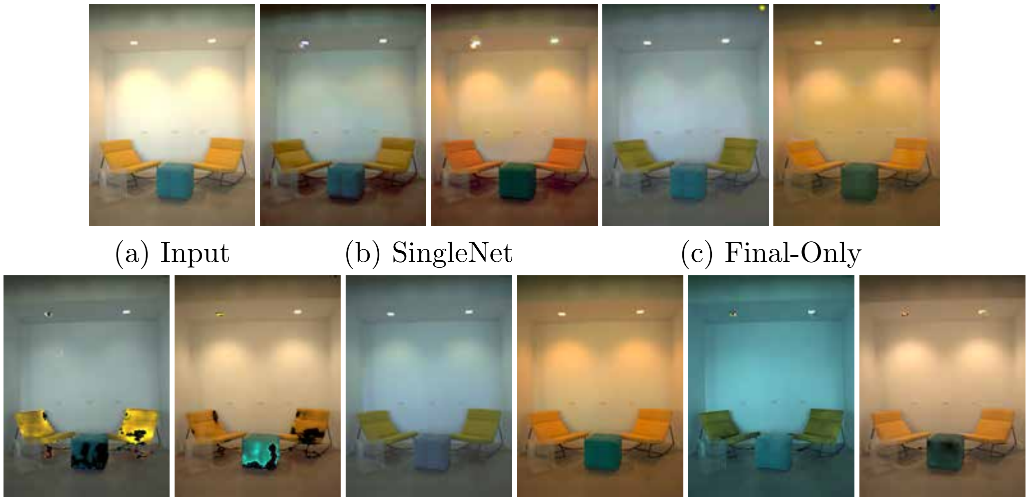
† Use additional information as input.

Table 5.1: We measure performance of versions of our network—trained with different kinds of supervision, and with different approaches to perform separation—as well as other baselines. Reported here are ℓ_1 error values for both estimated reflectance chromaticity (when available), as well as the final separated images.

evaluate the performance of the separated results, we compute the error for the separated result against the ground truth as:

$$\text{Loss} = \min\{E_{I_{1,1}} + E_{I_{2,2}}, E_{I_{1,2}} + E_{I_{2,1}}\} \tag{5.12}$$

where E denote the ℓ_1 error between two images. We use a global scale-invariant loss because we are most interested in capturing relative variations between the two images.



(a) Input (b) SingleNet (c) Final-Only

(d) Chrom-Only (e) Full-Direct (f) Full+ [81]

Figure 5.4: Qualitative comparison of image separation results of different versions of our network, as well as of the single encoder-decoder architecture network (SingleNet). We see that both SingleNet and our Final-only model both fail to separate the effects of illuminant shading from the input. Our Chrom-Only model yields a better result, but has severe artifacts in certain regions—highlighting that better chromaticity estimates do not lead to better separation. The results from our model with Full supervision yields the best results—with better separation of shadow and shading effects when we use its chromaticity outputs in conjunction with [81].

5.4.2 Quantitative results on synthetic benchmark

We next measure performance quantitatively on the synthetic dataset for our approach and compare it to several baselines and report these in Table 5.1. We begin by quantifying the importance of supervision. We train different models for our network: with full supervision, with supervision only on the quality of the final separated images (**Final-Only**), and training only the first sub-network, i.e., ChromNet, with supervision only on reflectance chromaticities (**Chrom-Only**). Moreover, for our fully supervised model (**Full**), we consider using the separated images directly predicted by our full network (**Full-Direct**), as well as taking only the reflectance chromaticity estimates and using Hui et al.'s algorithm [81]—which includes more complex processing—to perform separation (**Full+ [81]**). For the model with only chromaticity supervision, we use [81] as well to perform separation, and for the final-only supervised model (where intermediate chromaticities are not meaningful), we only consider the final output.

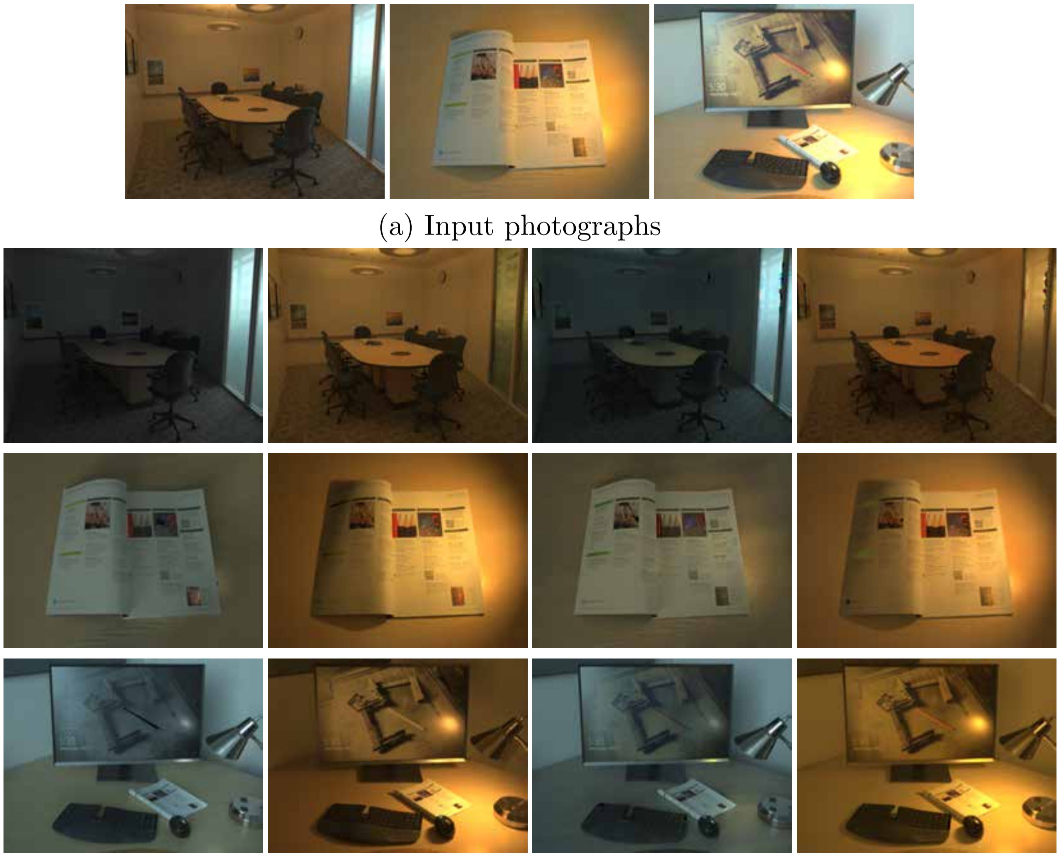
We find that our model trained with full supervision has the best performance in terms of the quality of final separated images. Interestingly, the **Chrom-Only** model is better at predicting chromaticity, but as expected, this does not translate to higher quality image outputs. The **Final-Only** model also yields worse separation results despite being trained with respect to their quality, highlighting the importance of intermediate supervision. Finally, we find that using our **Full** model in combination with [81] yields comparatively better results than taking the direct final output of the network. Thus, our final sub-networks (ShadingNet and SeparateNet) are able to only approximate [81]'s algorithm. Thus, their main benefit in our framework is in allowing back-propagation to provide supervision for chromaticity estimation, in a manner that is optimal for separation.

We also include comparisons to a network with a more traditional architecture (rather than three sub-networks) to do direct separation (**SingleNet**). We use the same architecture as the encoder-decoder portion of our ShadingNet, and train this again with supervision only on the final separated outputs. We find that this performs significantly worse (than even **Final-Only**), illustrating the utility of our physically-motivated architecture. Finally, we also include the comparisons with baselines where different intrinsic image decomposition methods [20, 95, 135] are used to estimate reflectance chromaticity from a single image, and these are used for separation with [81]. We find these methods yield lower accuracy in both reflectance chromaticity estimation and lighting separation—likely because they, like most intrinsic image methods, assume a single light source.

Finally, we evaluate on two methods that require additional information beyond a single image: ground truth light colors for Hsu et al. [77], and a flash/no-flash pair which provides direct access to reflectance chromaticity, for Hui et al. [81]. We produce better results than [77], but as expected, [81] yields the most accurate separation with direct access to chromaticity information — but requires capturing an additional flash image.

5.4.3 Qualitative evaluation on real data

Figure 5.4 shows results on a real image for the different versions of our network (as well as of **SingleNet**), while Figure 5.5 compares our results to Hui et al.'s method [81] when using a flash/no-flash pair. These results confirm our conclusions from Table 5.1 — we find that the version of our network trained with full supervision performs best, especially when used in combination with [81] to carry out the separation from predicted chro-



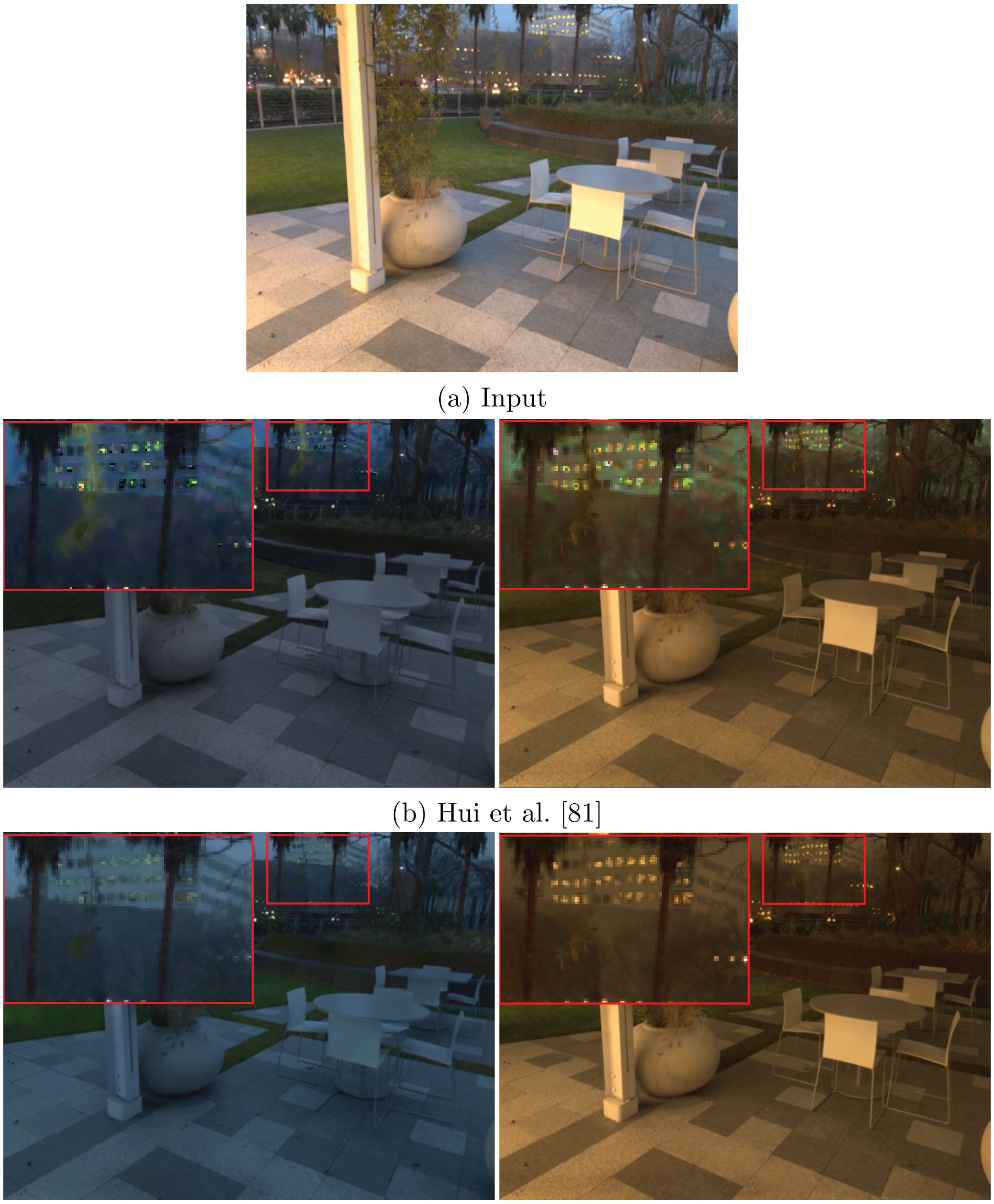
(a) Input photographs

(b) Hui et al. [81]

(c) Ours

Figure 5.5: We evaluate our technique against the flash photography technique by Hui et al [81]. While the proposed method may lead to small artifacts in the resulting image, we can achieve nearly the same visual quality as Hui et al. [81], which captures two photographs for the same scene. In comparison, the proposed technique by using single photograph yields more practical solution to the problem.

maticities. Moreover, despite requiring only a single image input, it comes close to matching Hui et al.'s [81] performance with a flash/no-flash pair. We show an example in Figure 5.6 where our method affords a distinct advantage even when an image with flash is available, but when several regions in the scene are too far from the flash. This leads to artifacts in those regions for [81], while our approach is able to perform a higher quality separation. The accompanying supplementary material contains additional results and comparisons for time-lapse videos and indoor scenes.



(a) Input

(b) Hui et al. [81]

(c) Ours

Figure 5.6: For the outdoor scene (a), the flash is not strong enough to illuminate the far-away scene points, which results in the artifacts in (b). In contrast, our method takes a single photograph and does not rely on flash illumination. As can be seen, the artifacts can be eliminated and visual quality has been significantly improved.

Chapter 6

Conclusions

In the dissertation, we present the approach to solve for the reflectance, shape and illumination of the scene from photographs. In particular, we demonstrate the feasibility of shape and SV-BRDFs estimation using the setup of photometric stereo, as well as the reflectance capture using a collocated light source and camera, a hardware setup that is commonly found in mobile devices. In addition, we have addressed the illumination analysis problem in terms of color constancy and light source separation. This ability to analyze and isolate lights in turn leads to state-of-the-art results on white balancing, intrinsic image decomposition, illumination editing, and color photometric stereo. We believe that this is a significant step towards true post-capture lighting control over images.

For the shape and reflectance estimation, we first present a photometric stereo technique for per-pixel normal and BRDF estimation for objects that are visually complex. We demonstrate that the use of a BRDF dictionary significantly simplifies the inverse problem and provides not just state-of-the-art results in normal and BRDF estimation but also works robustly on a wide range of real scenes. The hallmark of our approach is the ability to obtain surface normal and SV-BRDF estimates without requiring complex iterative techniques endemic to state-of-the-art techniques. Finally, our per-pixel estimation framework is ripe for further speed-ups by solve for the shape and reflectance at each pixel in parallel. To solve for the problem under the easy-to-deploy capture setup, we show that univariate sampling, commonly believed to be undesirable for reflectance estimation, can offer high-quality estimates of SV-BRDFs.

To estimate the illumination of the scene, we address two subset of problems. We first address the under-constrained problem of color constancy under complex spatially-varying illumination, and shown that using flash photography results in a closed-form solution to this problem. Our technique is automatic and does not rely on assumptions about the scene lighting or user inputs, which are endemic to all previous works. We further extend the scope of the problem by automatically separating an image into constituent images lit by each illuminant. This separation can be used to support applications like white balancing, lighting editing, intrinsic image decomposition, and RGB photometric stereo, where we demonstrate results that outperform state-of-the-art techniques on a wide range of images. However, the idea is build upon the use of a pair of flash and no-flash images, which is prone to the flash shadows, flash highlights as well as the movement between the image pair. To solve that, we train a deep neural network to predict the per-pixel reflectance chromaticity of the scene, which we use in conjunction with a previous flash/no-flash image-based separation algorithm to produce the final two output images.

6.1 Limitations

While our approach is good predicting at reflectance, geometry and scene illumination via the use of mobile device, it has issues regarding to practical applications concerns. We have detailed the limitations of the proposed methods as follows.

Calibration. To recover the shape and BRDF via the use of photometric setup, we require light calibration and hence, our method is most suited to shape and reflectance acquisition from light-stages where the light sources are fixed and the calibration is a one-time effort. To capture the reflectance via the mobile device, our method is limited to near-planar samples with little depth variation. This is because we rely on a planar geometric proxy to align the multiple captured images. The light intensity across the material sample should be uniform and significantly greater than the ambient light levels. Our method requires alignment for the input sequence. Imprecise alignment may lead to the blurry of the results

To estimate the illuminant colors and even separate the light sources, we assume that the scene is Lambertian, and hence our methods will fail on opaque objects that are extremely shiny (like mirrors). However, the incorrect results will be localized to the objects since the processing is largely per-pixel and the conic hull processing is inherently robust to outliers via the use of RANSAC and other pre-processing techniques.

For both cases, we assume the radiometric response of the camera is linear.

BRDF estimation. To estimate the per-pixel BRDF, we assume that the scene lies in the linear span of our dictionary. In the failure of this, our results can be unpredictable. Here, the need for a larger dictionary encompassing hundreds, if not thousands, of materials would be invaluable for the broader applicability of our method.

Illumination estimation. Our separation technique may fail to identify the correct illumination if there are no shadows in the scenes. A true planar scene with even two light sources can produce poor results in terms of source separation. Our experience has been that while separated images and illuminant colors are estimated incorrectly, relighting the scene often looks visually pleasing (even if non-realistic). While we enable the capability in lighting separation with single photograph by the use of deep neutral network, it is inherently limited by the data we used to train the network, where we require the input images are only illuminated by two illuminants.

6.2 Future work

To address the limitations and further improve the proposed method, we would like to continue working on the shape, reflectance and illumination estimation with respect to the following concerns.

Shape and reflectance estimation. Our method requires the calibration for both the light and camera, as well as the little depth variation of the objects, both of which make the method inconvenient and often difficult to use. Recently, we notice that mobile phone manufacturers have started producing a time-of-flight or structured-light direct depth sensor to their phones. This directly provides the depth measurements, and hence allows us to use it in aligning multiple captured images. In addition, by assuming the single point light source and collocated setup of camera and light source, we are able to solve for the lighting directions and camera poses. Now, we can apply our technique to solve for the reflectance and shape. This enables us to not only obviate the needs for the calibration, but also generalize our approach to the objects with arbitrary complex shape with SV-BRDFs.

More recently, deep neural network-based techniques have been proposed for estimating SV-BRDFs and shape of the objects from a single image [97]. While the method provides high-quality results for complex objects, it cannot be well generalized to the scenes with large distance to the camera by the use of flash light. Our long-term goal is to be able to recover the geometry of the scene as well as the reflectance of each object. Compared to past techniques in the field of intrinsic image, we focus on the objects with non-Lambertian BRDF and aim for introducing compact representation for BRDFs.

Reflectance and illumination editing. Given the separated image under single light source, a potential idea would be to use these estimates as a prior to guide the reflectance and shading separation. In that sense, it might be particularly interesting to incorporate the lighting separation to improve the intrinsic image decomposition. Note that the separated images share the same reflectance while being illuminated by different light source. This naturally provides a mini video sequence under a static camera, i.e. a two-image sequence for two light sources in the scene. We can initialize the reflectance and shading as the results generated by the method of [151], and introduce an iterative optimization scheme to refine the initial estimates. Another solution would be to feed the separated images into an extra sub-network to produce the reflectance and illuminant shadings.

Photometric stereo. Photometric stereo [152] seeks to estimate the shape of an object from images obtained from a static camera and under varying lighting. While it is able to return high quality estimates by the use of static camera, it requires multiple images as well as single light source in the scene. To reduce the number of input images, techniques have been proposed with a single shot where the object is lit by three monochromatic red, green, and blue, directional light sources [32]. However this estimation is still ill-posed and requires additional priors. Our technique by separating the illumination with respect to the light color provides a possibility to solve this problem. Similar to our method in two-light separation, we can pre-train a network to predict the reflectance chromaticity given the input images, which are illuminated by monochromatic red, green, and blue, directional light sources.

Anohter interesting application of our light source separation is in outdoor photometric stereo, which it is often necessary to produce the geometry of the buildings from a recorded time-lapse video. However, the outdoor illumination is not single directional light, which consists of both direct sunlight and indirect skylight. Given our technique to separate the sunlight from the skylight from a single photograph, we are able to produce the images only illuminated by the directional light source, i.e. sunlight. This enables us to apply Lambertian photometric stereo to generate the shape of the target buildings.

Relighting the scene. While our method is able to separate the light sources, the proposed method cannot estimate the position of the light source and relight the scene with novel illuminants. Recently, Gardner et al. [59] utilize the neutral network to solve the problem from LDR input images. Inspired by this work, the problem might be solved by adding an extra subnetwork to predict the location of the light source from our relative shading estimates. In contrast to the work of [59], we are able to remove the effects of reflectance and provide the shadings, which are significantly easier to learn for the position of light sources. To this end, the estimation might be explicitly enhanced to improve performance by using the relative shadings .

Computation power of mobile devices. While deep learning based techniques have received bulk of attention, the advances are not necessarily making networks more efficient with respect to size and speed. To apply proposed techniques in mobile devices, it is desirable to achieve good performance in a timely fashion on a computationally limited platform. To this end, we would like to continuously work along this line by developing efficient network architecture which is able to utilize our proposed physical constraint to reduce the complexity in the network design.

Bibliography

[1] Decoding raw digital photos in linux. *URL: https://www.cybercom.net/ dcoffin/dcraw/.*

[2] Light spectral power distribution database. *URL: lspdd.com/.*

[3] J. Ackermann, F. Langguth, S. Fuhrmann, A. Kuijper, and M. Goesele. Multi-view photometric stereo by example. In *3DV*, 2014.

[4] J. Ackermann, M. Ritz, A. Stork, and M. Goesele. Removing the example from example-based photometric stereo. In *Trends and Topics in Computer Vision*, pages 197–210. 2010.

[5] A. Agrawal, R. Chellappa, and R. Raskar. An algebraic approach to surface reconstruction from gradient fields. In *International Conference on Computer Vision (ICCV)*, 2005.

[6] M. Aittala, T. Aila, and J. Lehtinen. Reflectance modeling by neural texture synthesis. *ACM Transactions on Graphics (TOG)*, 35(4), 2016.

[7] M. Aittala, T. Weyrich, and J. Lehtinen. Two-shot svbrdf capture for stationary materials. *ACM Transactions on Graphics (TOG)*, 34(4):110, 2015.

[8] Y. Aksoy, C. Kim, P. Kellnhofer, S. Paris, M. Elgharib, M. Pollefeys, and W. Matusik. A dataset of flash and ambient illumination pairs from the crowd. In *ECCV*, 2018.

[9] N. Alldrin, T. Zickler, and D. Kriegman. Photometric stereo with non-parametric and spatially-varying reflectance. In *Computer Vision and Pattern Recognition (CVPR)*, 2008.

[10] N. G. Alldrin and D. Kriegman. Toward reconstructing surfaces with arbitrary isotropic reflectance: A stratified photometric stereo approach. In *ICCV*, 2007.

[11] N. G. Alldrin, S. P. Mallick, and D. J. Kriegman. Resolving the generalized bas-relief ambiguity by entropy minimization. In *CVPR*, 2007.

[12] E. M. Arkin, Y.-J. Chiang, M. Held, J. S. B. Mitchell, V. Sacristan, S. Skiena, and T.-C. Yang. On minimum-area hulls. *Algorithmica*, 21(1):119–136, 1998.

[13] M. Ashikhmin and P. Shirley. An anisotropic phong brdf model. *Journal of graphics tools*, 5(2):25–32, 2000.

[14] J. T. Barron. Convolutional color constancy. 2015.

[15] J. T. Barron and J. Malik. Color constancy, intrinsic images, and shape estimation. In *ECCV*. 2012.

[16] J. T. Barron and J. Malik. Shape, albedo, and illumination from a single image of an unknown object. In *CVPR*, 2012.

[17] J. T. Barron and J. Malik. Shape, illumination, and reflectance from shading. *PAMI*, 37(8):1670–1687, 2015.

[18] H. Barrow and J. Tenenbaum. Recovering intrinsic scene characteristics from images. *Computer Vision Systems*, 1978.

[19] R. Basri and D. Jacobs. Lambertian reflectance and linear subspaces. *PAMI*, 25(2):218–233, 2003.

[20] S. Bell, K. Bala, and N. Snavely. Intrinsic images in the wild. *TOG*, 33(4):159, 2014.

[21] B. Bitterli. Rendering resources, 2016. https://benedikt-bitterli.me/resources/.

[22] J. F. Blinn and M. E. Newell. Texture and reflection in computer generated images. *Communications of the ACM*, 19(10):542–547, 1976.

[23] N. Bonneel, B. Kovacs, S. Paris, and K. Bala. Intrinsic decompositions for image editing. In *Computer Graphics Forum*, 2017.

[24] N. Bonneel, K. Sunkavalli, J. Tompkin, D. Sun, S. Paris, and H. Pfister. Interactive intrinsic video editing. *TOG*, 33(6):197, 2014.

[25] A. Bousseau, S. Paris, and F. Durand. User-assisted intrinsic images. In *TOG*, volume 28, page 130, 2009.

[26] I. Boyadzhiev, K. Bala, S. Paris, and F. Durand. User-guided white balance for mixed lighting conditions. *TOG*, 31(6):200, 2012.

[27] I. Boyadzhiev, S. Paris, and K. Bala. User-assisted image compositing for photographic lighting. *TOG*, 32(4):36–1, 2013.

[28] G. J. Brostow, C. Hernández, G. Vogiatzis, B. Stenger, and R. Cipolla. Video normals from colored lights. *PAMI*, 33(10):2104–2114, 2011.

[29] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *J. ACM*, 58(3):11, 2011.

[30] R. Carroll, R. Ramamoorthi, and M. Agrawala. Illumination decomposition for material recoloring with consistent interreflections. *TOG*, 30(4):43, 2011.

[31] A. Chakrabarti. Color constancy by learning to predict chromaticity from luminance. In *NIPS*, 2015.

[32] A. Chakrabarti and K. Sunkavalli. Single-image rgb photometric stereo with spatially-varying albedo. In *3DV*, 2016.

[33] M. Chandraker. On joint shape and material recovery from motion cues. In *European Conference on Computer Vision (ECCV)*, 2014.

[34] M. Chandraker. What camera motion reveals about shape with unknown brdf. In *Computer Vision and Pattern Recognition (CVPR)*, 2014.

[35] M. Chandraker. The information available to a moving observer on shape with unknown, isotropic brdfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 38(7):1283–1297, 2016.

[36] M. Chandraker, J. Bai, and R. Ramamoorthi. On differential photometric reconstruction for unknown, isotropic brdfs. *PAMI*, 35(12):2941–2955, 2013.

[37] M. Chandraker and R. Ramamoorthi. What an image reveals about material reflectance. In *ICCV*, 2011.

[38] H. Y. Chong, S. J. Gortler, and T. Zickler. The von kries hypothesis and a basis for color constancy. In *ICCV*, 2007.

[39] R. L. Cook and K. E. Torrance. A reflectance model for computer graphics. *ACM Transactions on Graphics (TOG)*, 1(1):7–24, 1982.

[40] K. J. Dana, B. van Ginneken, S. K. Nayar, and J. J. Koenderink. Reflectance and texture of real-world surfaces. *ACM Transactions on Graphics (TOG)*, 18(1):1–34, 1999.

[41] P. Debevec. Rendering synthetic objects into real scenes: Bridging traditional and image-based graphics with global illumination and high dynamic range photography. In *ACM SIGGRAPH classes*, 2008.

[42] P. Debevec. The Light Stages and Their Applications to Photoreal Digital Actors. In *SIGGRAPH Asia*, 2012.

[43] P. Debevec, T. Hawkins, C. Tchou, H.-P. Duiker, W. Sarokin, and M. Sagar. Acquiring the reflectance field of a human face. In *Computer Graphics and Interactive Techniques*, pages 145–156, 2000.

[44] M. S. Drew, H. R. V. Joze, and G. D. Finlayson. Specularity, the zeta-image, and information-theoretic illuminant estimation. In *ECCV*, pages 411–420, 2012.

[45] M. Ebner. Color constancy using local color shifts. In *ECCV 2004*. 2004.

[46] P. Einarsson, C.-F. Chabert, A. Jones, W.-C. Ma, B. Lamond, T. Hawkins, M. Bolas, S. Sylwan, and P. Debevec. Relighting human locomotion with flowed reflectance fields. In *Eurographics Conference on Rendering Techniques*, pages 183–194, 2006.

[47] P. Einarsson, C.-F. Chabert, A. Jones, W.-C. Ma, B. Lamond, T. Hawkins, M. T. Bolas, S. Sylwan, and P. E. Debevec. Relighting human locomotion with flowed reflectance fields. In *Rendering techniques*, page 17, 2006.

[48] E. Eisemann and F. Durand. Flash photography enhancement via intrinsic relighting. In *TOG*, volume 23, pages 673–678, 2004.

[49] D. Eppstein, M. Overmars, G. Rote, and G. Woeginger. Finding minimum areak-gons. *Discrete & Computational Geometry*, 7(1):45–58, 1992.

[50] P. Favaro and T. Papadhimitri. A closed-form solution to uncalibrated photometric stereo via diffuse maxima. In *CVPR*, 2012.

[51] M. Fazel, H. Hindi, and S. P. Boyd. A rank minimization heuristic with application to minimum order system approximation. In *American Control Conference*, volume 6, pages 4734–4739, 2001.

[52] G. Finlayson, M. Drew, and B. Funt. Enhancing von kries adaptation via sensor transformations. 1993.

[53] G. Finlayson and S. Hordley. Improving gamut mapping color constancy. *TIP*, 9(10):1774–1783, 2000.

[54] G. D. Finlayson. Color constancy in diagonal chromaticity space. In *ICCV*, 1995.

[55] G. D. Finlayson, M. S. Drew, and B. V. Funt. Diagonal transforms suffice for color constancy. In *ICCV*, 1993.

[56] Y. Furukawa, C. Hernández, et al. Multi-view stereo: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 9(1-2):1–148, 2015.

[57] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. *PAMI*, 32(8):1362–1376, 2010.

[58] G. Fyffe, X. Yu, and P. Debevec. Single-shot photometric stereo by spectral multiplexing. In *ICCP*, 2011.

[59] M.-A. Gardner, K. Sunkavalli, E. Yumer, X. Shen, E. Gambaretto, C. Gagn, and J.-F. Lalonde. Learning to predict indoor illumination from a single image. In *TOG*, 2017.

[60] P. V. Gehler, C. Rother, A. Blake, T. Minka, and T. Sharp. Bayesian color constancy revisited. In *CVPR*, 2008.

[61] A. Gijsenij, T. Gevers, and J. van de Weijer. Computational color constancy: Survey and experiments. *TIP*, 20(9):2475–2489, 2011.

[62] A. Gijsenij, R. Lu, and T. Gevers. Color constancy for multiple light sources. *TIP*, 21(2):697–707, 2012.

[63] D. B. Goldman, B. Curless, A. Hertzmann, and S. Seitz. Shape and spatially-varying BRDFs from photometric stereo. In *ICCV*, 2005.

[64] D. B. Goldman, B. Curless, A. Hertzmann, and S. Seitz. Shape and spatially-varying brdfs from photometric stereo. In *International Conference on Computer Vision (ICCV)*, 2005.

[65] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 2.1. `http://cvxr.com/cvx`, 2014.

[66] R. Grosse, M. K. Johnson, E. H. Adelson, and W. T. Freeman. Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *CVPR*, 2009.

[67] R. Harman and V. Lacko. On decompositional algorithms for uniform sampling from n-spheres and n-balls. *Journal of Multivariate Analysis*, 101(10):2297–2304, 2010.

[68] D. Hauagge, S. Wehrwein, K. Bala, and N. Snavely. Photometric ambient occlusion. In *CVPR*, 2013.

[69] M. Hauta-Kasari, K. Miyazawa, S. Toyooka, and J. Parkkinen. Spectral vision system for measuring color images. *JOSA A*, 16(10):2352–2362, 1999.

[70] X. D. He, K. E. Torrance, F. X. Sillion, and D. P. Greenberg. A comprehensive physical model for light reflection. *TOG*, 25(4):175–186, 1991.

[71] A. Hertzmann and S. Seitz. Example-based photometric stereo: Shape reconstruction with general, varying BRDFs. *PAMI*, 27(8):1254–1264, 2005.

[72] T. Higo, Y. Matsushita, and K. Ikeuchi. Consensus photometric stereo. In *CVPR*, 2010.

[73] T. Higo, Y. Matsushita, N. Joshi, and K. Ikeuchi. A hand-held photometric stereo camera for 3-d modeling. In *International Conference on Computer Vision (ICCV)*, 2009.

[74] Y. Hold-Geoffroy, K. Sunkavalli, S. Hadap, E. Gambaretto, and J.-F. Lalonde. Deep outdoor illumination estimation. In *CVPR*, 2017.

[75] Y. Hold-Geoffroy, K. Sunkavalli, S. Hadap, E. Gambaretto, and J.-F. Lalonde. Deep outdoor illumination estimation. In *CVPR*, 2017.

[76] B. K. Horn. *Obtaining shape from shading information.* 1989.

[77] E. Hsu, T. Mertens, S. Paris, S. Avidan, and F. Durand. Light mixture estimation for spatially varying white balance. In *TOG*, volume 27, page 70, 2008.

[78] Z. Hui and A. Sankaranarayanan. Shape and spatially-varying reflectance estimation from virtual exemplars. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 1(99):1, 2016.

[79] Z. Hui and A. C. Sankaranarayanan. A dictionary-based approach for estimating shape and spatially-varying reflectance. In *International Conference on Computational Photography (ICCP)*, 2015.

[80] Z. Hui, A. C. Sankaranarayanan, K. Sunkavalli, and S. Hadap. White balance under mixed illumination using flash photography. In *ICCP*, 2016.

[81] Z. Hui, K. Sunkavalli, S. Hadap, and A. C. Sankaranarayanan. Illuminant spectra-based source separation using flash photography. In *CVPR*, 2018.

[82] S. Ikehata and K. Aizawa. Photometric stereo using constrained bivariate regression for general isotropic surfaces. In *CVPR*, 2014.

[83] S. Ikehata, D. Wipf, Y. Matsushita, and K. Aizawa. Robust photometric stereo using sparse regression. In *CVPR*, 2012.

[84] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017.

[85] W. Jakob. Mitsuba renderer, 2010. *URL: http://www. mitsuba-renderer. org*, 3:10, 2015.

[86] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.

[87] M. K. Johnson and E. H. Adelson. Shape estimation in natural illumination. In *CVPR*, 2011.

[88] H. R. V. Joze and M. S. Drew. Exemplar-based color constancy and multiple illumination. *PAMI*, 36(5):860–873, 2014.

[89] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[90] P.-Y. Laffont and J.-C. Bazin. Intrinsic decomposition of image sequences from local temporal variations. In *ICCV*, 2015.

[91] E. P. Lafortune, S.-C. Foo, K. E. Torrance, and D. P. Greenberg. Non-linear approximation of reflectance functions. In *Computer graphics and interactive techniques*, 1997.

[92] J.-F. Lalonde, S. G. Narasimhan, and A. A. Efros. What do the sun and the sky tell us about the camera? *IJCV*, 88(1):24–51, 2010.

[93] J. Lawrence, A. Ben-Artzi, C. DeCoro, W. Matusik, H. Pfister, R. Ramamoorthi, and S. Rusinkiewicz. Inverse shade trees for non-parametric material representation and editing. *ACM Transactions on Graphics (TOG)*, 25(3):735–745, 2006.

[94] Y. Li and M. S. Brown. Single image layer separation using relative smoothness. In *CVPR*, 2014.

[95] Z. Li and N. Snavely. Cgintrinsics: Better intrinsic image decomposition through physically-based rendering. In *ECCV*, 2018.

[96] Z. Li, K. Sunkavalli, and M. Chandraker. Materials for masses: Svbrdf acquisition with a single mobile phone image. In *ECCV*, 2018.

[97] Z. Li, Z. Xu, R. Ramamoorthi, K. Sunkavalli, and M. Chandraker. Learning to reconstruct shape and spatially-varying reflectance from a single image. In *TOG*, page 269, 2018.

[98] D. Lischinski, Z. Farbman, M. Uyttendaele, and R. Szeliski. Interactive local adjustment of tonal values. *TOG*, 25(3):646–653, 2006.

[99] F. Lu, Y. Matsushita, I. Sato, T. Okabe, and Y. Sato. Uncalibrated photometric stereo for unknown isotropic reflectances. In *CVPR*, 2013.

[100] F. Lu, Y. Matsushita, I. Sato, T. Okabe, and Y. Sato. From intensity profile to surface normal: photometric stereo for unknown light sources and isotropic reflectances. *PAMI*, 37:1999–2012, 2015.

[101] F. Lu, I. Sato, and Y. Sato. Uncalibrated photometric stereo based on elevation angle recovery from brdf symmetry of isotropic materials. In *CVPR*, 2015.

[102] S. R. Marschner, S. H. Westin, E. P. Lafortune, K. E. Torrance, and D. P. Greenberg. Image-based brdf measurement including human skin. In *Rendering Techniques*. 1999.

[103] W. Matusik, H. Pfister, M. Brand, and L. McMillan. A data-driven reflectance model. *ACM Transactions on Graphics (TOG)*, 22(3):759–769, 2003.

[104] A. Medvedeva and A. Mukhopadhyay. An implementation of a linear time algorithm for computing the minimum perimeter triangle enclosing a convex polygon. In *CCCG*, volume 3, pages 25–28, 2003.

[105] S. G. Narasimhan, S. K. Nayar, B. Sun, and S. J. Koppal. Structured light in scattering media. In *ICCV*, 2005.

[106] S. K. Nayar, G. Krishnan, M. D. Grossberg, and R. Raskar. Fast separation of direct and global components of a scene using high frequency illumination. In *TOG*, volume 25, pages 935–944, 2006.

[107] A. Ngan, F. Durand, and W. Matusik. Experimental analysis of brdf models. *Rendering Techniques*, 2005(16):2, 2005.

[108] J. B. Nielsen, H. W. Jensen, and R. Ramamoorthi. On optimal, minimal brdf sampling for reflectance acquisition. *ACM Transactions on Graphics (TOG)*, 34(6):186, 2015.

[109] M. Oren and S. K. Nayar. Generalization of the lambertian model and implications for machine vision. *International Journal on Computer Vision (IJCV)*, 14(3):227–251, 1995.

[110] J. O'Rourke, A. Aggarwal, S. Maddila, and M. Baldwin. An optimal algorithm for finding minimal enclosing triangles. *Journal of Algorithms*, 7(2):258–269, 1986.

[111] G. Oxholm and K. Nishino. Shape and reflectance from natural illumination. In *European Conference on Computer Vision (ECCV)*, 2012.

[112] G. Oxholm and K. Nishino. Multiview shape and reflectance from natural illumination. In *Computer Vision and Pattern Recognition (CVPR)*, 2014.

[113] G. Oxholm and K. Nishino. Shape and reflectance estimation in the wild. *PAMI*, 38(2):376–389, 2016.

[114] G. Oxholm and K. Nishino. Shape and reflectance estimation in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 38(2):376–389, 2016.

[115] T. Papadhimitri and P. Favaro. A new perspective on uncalibrated photometric stereo. In *CVPR*, 2013.

[116] N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3):123–231, 2013.

[117] O. Pârvu and D. Gilbert. Implementation of linear minimum area enclosing triangle algorithm. *Computational and Applied Mathematics*, pages 1–16, 2014.

[118] P. Peers, N. Tamura, W. Matusik, and P. Debevec. Post-production facial performance relighting using reflectance transfer. *TOG*, 26(3), 2007.

[119] G. Petschnigg, R. Szeliski, M. Agrawala, M. Cohen, H. Hoppe, and K. Toyama. Digital photography with flash and no-flash image pairs. *TOG*, 23(3):664–672, 2004.

[120] V. Prinet, D. Lischinski, and M. Werman. Illuminant chromaticity from image sequences. In *ICCV*, 2013.

[121] R. Ramamoorthi. Analytic PCA construction for theoretical analysis of lighting variability in images of a Lambertian object. *PAMI*, 24(10):1322–1333, 2002.

[122] R. Raskar, K.-H. Tan, R. Feris, J. Yu, and M. Turk. Non-photorealistic camera: depth edge detection and stylized rendering using multi-flash imaging. In *TOG*, volume 23, pages 679–688, 2004.

[123] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

[124] K. Rematas, T. Ritschel, M. Fritz, E. Gavves, and T. Tuytelaars. Deep reflectance maps. In *CVPR*, 2016.

[125] P. Ren, J. Wang, J. Snyder, X. Tong, and B. Guo. Pocket reflectometry. 30(4):45, 2011.

[126] C. Riess, E. Eibenberger, and E. Angelopoulou. Illuminant color estimation for real-world mixed-illuminant scenes. In *ICCVW*, 2011.

[127] J. Riviere, P. Peers, and A. Ghosh. Mobile surface reflectometry. In *Computer Graphics Forum*, 2016.

[128] F. Romeiro, Y. Vasilyev, and T. Zickler. Passive reflectometry. In *ECCV*, 2008.

[129] F. Romeiro, Y. Vasilyev, and T. Zickler. Passive reflectometry. In *European Conference on Computer Vision (ECCV)*, 2008.

[130] F. Romeiro and T. Zickler. Blind reflectometry. In *ECCV*, 2010.

[131] F. Romeiro and T. Zickler. Blind reflectometry. In *European Conference on Computer Vision (ECCV)*, 2010.

[132] S. Rusinkiewicz. A new change of variables for efficient brdf representation. In *Eurographics Workshop on Rendering*, 1998.

[133] S. M. Rusinkiewicz. A new change of variables for efficient brdf representation. In *Rendering techniques*, pages 11–22. 1998.

[134] J. Shen, X. Yang, Y. Jia, and X. Li. Intrinsic images using optimization. In *CVPR*, 2011.

[135] J. Shen, X. Yang, X. Li, and Y. Jia. Intrinsic image decomposition using optimization and user scribbles. *IEEE Transactions on Cybernetics*, 43(2):425–436, 2013.

[136] B. Shi, P. Tan, Y. Matsushita, and K. Ikeuchi. Elevation angle from reflectance monotonicity: Photometric stereo for general isotropic reflectances. In *ECCV*. 2012.

[137] B. Shi, P. Tan, Y. Matsushita, and K. Ikeuchi. Bi-polynomial modeling of low-frequency reflectances. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, 36(6):1078–1091, 2014.

[138] B. Shi, Z. Wu, Z. Mo, D. Duan, and S.-K. Y. P. Tan. A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo. In *CVPR*, 2016.

[139] R. Shiradkar, L. Shen, G. Landon, S. Heng Ong, and P. Tan. A new perspective on material classification and ink identification. In *Computer Vision and Pattern Recognition (CVPR)*, 2014.

[140] W. M. Silver. *Determining shape and reflectance using multiple images.* PhD thesis, Massachusetts Institute of Technology, 1980.

[141] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser. Semantic scene completion from a single depth image. *CVPR*, 2017.

[142] J. Sun, J. Sun, S. B. Kang, Z.-B. Xu, J. Sun, and H.-Y. Shum. Flash cut: Foreground extraction with flash and no-flash image pairs. *CVPR*, 2007.

[143] J. Sun, J. Sun, S. B. Kang, Z.-B. Xu, X. Tang, and H.-Y. Shum. Flash cut: Foreground extraction with flash and no-flash image pairs. In *CVPR*, 2007.

[144] K. Sunkavalli, F. Romeiro, W. Matusik, T. Zickler, and H. Pfister. What do color changes reveal about an outdoor scene? In *CVPR*, 2008.

[145] K. Sunkavalli, T. Zickler, and H. Pfister. Visibility subspaces: Uncalibrated photometric stereo with shadows. In *ECCV*. 2010.

[146] P. Tan, L. Quan, and T. Zickler. The geometry of reflectance symmetries. *PAMI*, 33(12):2506–2520, 2011.

[147] R. T. Tan, K. Ikeuchi, and K. Nishino. Color constancy through inverse-intensity chromaticity space. In *Digitally Archiving Cultural Objects*, pages 323–351. 2008.

[148] Y. Tang, R. Salakhutdinov, and G. Hinton. Deep lambertian networks. *arXiv preprint arXiv:1206.6445*, 2012.

[149] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *IJCV*, 9(2):137–154, 1992.

[150] G. J. Ward. Measuring and modeling anisotropic reflection. *ACM Transactions on Graphics (TOG)*, 26(2):265–272, 1992.

[151] Y. Weiss. Deriving intrinsic images from image sequences. In *ICCV*, 2001.

[152] R. Woodham. Photometric method for determining surface orientation from multiple images. *Optical engineering*, 19(1):191139, 1980.

[153] R. J. Woodham. Photometric method for determining surface orientation from multiple images. *Optical Engineering*, 19(1), 1980.

[154] C. Wu. Towards linear-time incremental structure from motion. In *3DV*, 2013.

[155] L. Wu, A. Ganesh, B. Shi, Y. Matsushita, Y. Wang, and Y. Ma. Robust photometric stereo via low-rank matrix completion and recovery. In *ACCV*. 2011.

[156] Z. Xu, J. B. Nielsen, J. Yu, H. W. Jensen, and R. Ramamoorthi. Minimal brdf sampling for two-shot near-field reflectance acquisition. *ACM Transactions on Graphics (TOG)*, 35(6):188, 2016.

[157] C. Yu, Y. Seo, and S. W. Lee. Photometric stereo from maximum feasible Lambertian reflections. In *ECCV*. 2010.

[158] H. Zhang, K. Dana, and K. Nishino. Reflectance hashing for material recognition. In *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[159] Q. Zhao, P. Tan, Q. Dai, L. Shen, E. Wu, and S. Lin. A closed-form solution to retinex with nonlocal texture constraints. *PAMI*, 34(7):1437–1444, 2012.

[160] T. Zhou, P. Krahenbuhl, and A. A. Efros. Learning data-driven reflectance priors for intrinsic image decomposition. In *ICCV*, 2015.

[161] Z. Zhou, Z. Wu, and P. Tan. Multi-view photometric stereo with spatially varying isotropic materials. In *CVPR*, 2013.

[162] S. Zhuo, D. Guo, and T. Sim. Robust flash deblurring. In *CVPR*, 2010.