Live Inverse Rendering

A Dissertation submitted towards the degree
Doctor of Engineering (Dr.-Ing.)
of the Faculty of Mathematics and Computer Science
of Saarland University

Abhimitra Meka

Saarbrücken
2019

Dekan - Dean:

Prof. Dr. Sebastian Hack
Saarland University
Saarbrücken, Germany

Kolloquiums - Defense

Datum - Date
February 3, 2020, in Saarbrücken

Vorsitzender - Head of Colloquium:
Prof. Dr. Antonio Krüger

Prüfer - Examiners:

Prof. Dr. Christian Theobalt

Prof. Dr. George Drettakis

Dr. Michael Zollhöfer

Protokoll - Reporter:

Dr. Florian Bernard

To my parents, Kalyani (Suneeta) and Sudhakar

Acknowledgements

I am very grateful to the invaluable guidance of my thesis advisor, Prof. Christian Theobalt. I am constantly amazed by his consistent dedication to guiding his students through all the processes of research – from the details of research writing to deep-dives into mathematical models in computer graphics to the larger scale trends in scientific enquiry – while at the same time giving a free reign to our curiosities and motivations. I am also thankful to him for the freedom and accommodations for handling my personal obligations at various times during my doctoral research, balancing work and personal lives is a challenge for so many researchers, but it is a bit easier when you have a genuinely kind advisor.

I am also grateful to Prof. Hans-Peter Seidel and the institution of MPI Informatics, particularly the computer graphics department, for the open, innovative and collaborative research environment. My work here was funded by the ERC grants, CapReal and 4DReply, I am thankful to the European Research Consortium and Christian Theobalt for funding me.

I am very thankful to the administrative assistants in the department, Sabine Budde and Ellen Fries. The department would be quite lost without their support with the administrative processes, and also the warmth they bring to our often-times narrow research lives with the various events they organize.

I am deeply indebted to the two of the biggest influences on my doctoral research, my post-docs Michael Zollhöfer and Christian Richardt. Having their consistent guidance, particularly in the early stages of research, was crucial to my growth. They set the high standards for various research projects that we worked on together. To borrow words from Sir Isaac Newton, my work was facilitated by standing on the shoulders of these two giants.

I am thankful to Shahram Izadi's team at Google for the opportunity to collaborate during my internship. I thoroughly enjoyed working with Christoph Rhemann, Sean Fanello, Julien Valentin, Christian Häne, Rohit Pandey, Paul Debevec and the whole Lightstage team. It was

an invaluable experience in bringing together academic research and industry practice.

I am thankful to the post-docs in the department – Avishek Chatterjee, Mohamed Elgharib and Shida Beigpour – for the many long discussions and research ideas that came out of it. I also had an opportunity to work with incredible pre-doctoral students – Mohammad Shafiei, Maxim Maximov and Gereon Fox. The nature of our work involved things like shopping for hours for toys and lugging around camera equipment to strange places. We burnt a lot of midnight oil in our work, but their energy and enthusiasm made it a great deal of fun.

I owe so much to my peers, the fellow doctoral students in the department, whose zeal for their research has been a source of inspiration. Helge Rhodin, Pablo Garrido, Nadia Robertini, Hyeongwoo Kim, Franziska Müller, Ayush Tewari, Lingjie Liu, Petr Kellnhofer, Jozef Hladky, Marc Habermann, Ikhsanul Habibie, Jiayi Wang – I am thankful to them for the discussions over the years and also sharing this journey with me. I owe a special thanks to Srinath Sridhar and Dushyant Mehta who were not only motivational in their work, but also fellow residents of the bittersweet address of Gaußtraße 81 and (limited-time) gym and running buddies, their company ensured that my physical and mental health was taken care of to some extent! I am also thankful to my wonderful office-mates, Dan Casas and Edgar Tretschk, I learnt quite a bit about their research simply through osmosis.

The seeds of my interest in scientific research were sowed much before I came to MPI, by a wonderful set of teachers. I am very thankful to Prof. Subhasis Chaudhuri, my advisor for the Masters' thesis, it was his courses in image processing and computer vision that led to my interest in this field. My high school teachers – Mr. Gurwara, Mr. Seshu, Mrs. Savitha, Mrs. Narayanan, Ms. Irvinder and many more – instilled in me not only an interest in scientific practice, but also personal and social values that have been crucial in every endeavour that I have undertaken.

The motivation of any and all of my work is my family. I cannot put into words the support and sacrifice from their side that enabled this research work. I am forever in gratitude to my parents, Kalyani and Sudhakar, and my brother, Arun for this. I thank my family for being there for me in every which way throughout this journey.

Abstract

The field of computer graphics is being transformed by the process of 'personalization'. The advent of augmented and mixed reality technology is challenging the existing graphics systems, which traditionally required elaborate hardware and skilled artistic efforts. Now, photorealistic graphics are require to be rendered on mobile devices with minimal sensors and compute power, and integrated with the real world environment automatically. Seamlessly integrating graphics into real environments requires the estimation of the fundamental light transport components of a scene - geometry, reflectance and illumination. While estimating environmental geometry and self-localization on mobile devices has progressed rapidly, the task of estimating scene reflectance and illumination from monocular images or videos in real-time (termed *live inverse rendering*) is still at a nascent stage. The challenge is that of designing efficient representations and models for these appearance parameters and solving the resulting high-dimensional, non-linear and under-constrained system of equations at frame rate. This thesis comprehensively explores, for the first time, various representations, formulations, algorithms and systems for addressing these challenges in monocular inverse rendering. Starting with simple assumptions on the light transport model – of Lambertian surface reflectance and single light bounce scenario – the thesis expands in various directions by including 3D geometry, multiple light bounces, non-Lambertian isotropic surface reflectance and data-driven reflectance representation to address various facets of this problem. In the first part, the thesis explores the design of fast parallel non-linear GPU optimization schemes for solving both sparse and dense set of equations underlying the inverse rendering problem. In the next part, it applies the current advances in machine learning methods to design novel formulations and loss-energies to give a significant push to the state-of-the-art of reflectance and illumination estimation. Several real-time applications of illumination-aware scene editing, including relighting and material-cloning, are also shown to be made possible for first time by the new models proposed in this thesis. Finally, an outlook for future work on this problem is laid out, with particular emphasis on the interesting new opportunities afforded by the recent advances in machine learning.

Kurzfassung

Das Gebiet der Computergrafik wird durch den Prozess der Personalisierung verndert. Das Aufkommen von Augmented- und Mixed-Reality-Technologie stellt die bestehenden Grafiksysteme vor Herausforderungen, die bisher aufwndige Hardware und geschickte knstlerische Bemhungen erforderten. Fotorealistische Grafiken mssen jetzt auf Mobilgerten mit minimalem Sensor- und Rechenaufwand gerendert und automatisch in die reale Umgebung integriert werden. Die nahtlose Integration von Grafiken in reale Umgebungen erfordert die Abschtzung der grundlegenden Lichttransportkomponenten einer Szene: Geometrie, Reflektanz und Beleuchtung. Whrend die Schtzung der Umgebungsgeometrie und die Selbstlokalisierung auf Mobilgerten rasch Fortschritte gemacht hat, befindet sich die Schtzung der Reflektanz und der Beleuchtung von Szenen anhand von Monokularbildern oder -videos in Echtzeit (als Live-Inverse-Rendering bezeichnet) noch im Anfangsstadium. Die Herausforderung besteht darin, effiziente Darstellungen und Modelle fr diese Parameter zu entwerfen und das resultierende hochdimensionale, nichtlineare und unterspezifizierte Gleichungssystem in Echtzeit zu lsen. In dieser Arbeit werden erstmals verschiedene Darstellungen, Formulierungen, Algorithmen und Systeme zur Bewltigung dieser Herausforderungen beim monokularen Inverse-Rendering umfassend untersucht. Ausgehend von einfachen Annahmen zum Lichttransportmodell der Lambertschen Oberflchenreflektanz und Einfachlichtreflexionen geht diese Arbeit in mehrere Richtungen, indem 3D-Geometrie, Mehrfachlichtreflexionen, nicht-Lambertsche isotrope Oberflchenreflektanz und datengesttzte Reflektanzmodelle hinzugefgt und damit verschiedene Facetten des Problems bercksichtigt werden. Im ersten Teil der Arbeit wird der Entwurf von schnellen, parallelen, nichtlinearen GPU-Optimierungsschemata untersucht, mit denen sowohl dnnbesetzte als auch dichte Gleichungssysteme gelst werden knnen, die dem Inverse-Rendering-Problem zugrunde liegen. Im nchsten Teil werden aktuellen Fortschritte des maschinellen Lernens angewendet, um neuartige Formulierungen und Zielfunktionen zu entwerfen, die den Stand der Wissenschaft bei der Schtzung der Reflektanz und der Beleuchtung deutlich vorantreiben. Mit den in dieser Arbeit vorgeschlagenen neuen Modellen

werden auch erstmals Echtzeitanwendungen fr die Bearbeitung von Szenen unter Bercksichtigung der Beleuchtung, wie Beleuchtungsverndderung und Materialklonen, realisiert. Schlielich wird ein Ausblick fr die knftige Arbeit an diesem Problem gegeben, wobei der Schwerpunkt auf den interessanten neuen Mglichkeiten liegt, die sich aus den jngsten Fortschritten des maschinellen Lernens ergeben.

Contents

Nomenclature xix

1 Introduction 1
- 1.1 Motivation . . . 1
- 1.2 Overview . . . 4
- 1.3 Structure . . . 5
- 1.4 Contributions . . . 6

2 Technical Background 9
- 2.1 Image Formation . . . 9
  - 2.1.1 Rendering Equation . . . 9
  - 2.1.2 Inverse Rendering . . . 10
- 2.2 Reflectance . . . 11
  - 2.2.1 Reflectance models . . . 12
- 2.3 Illumination . . . 13
  - 2.3.1 Environment map representation . . . 14
  - 2.3.2 Spherical harmonics illumination . . . 14
- 2.4 Intrinsic Decomposition . . . 16

I Real-time Intrinsic Decomposition 19

3 Live Intrinsic Video 21
- 3.1 Introduction . . . 21
- 3.2 Related Work . . . 24
- 3.3 Overview . . . 27
- 3.4 Energy . . . 27
  - 3.4.1 Data Fitting Term . . . 28
  - 3.4.2 Local Prior Terms . . . 29

3.4.3 Spatio-Temporal Reflectance Consistency Prior 31
3.4.4 Reflectance Clustering Prior 32
3.5 Optimization 33
3.5.1 Data-Parallel IRLS Core Solver 34
3.5.2 LocalGlobal Optimization Approach 36
3.5.3 Adding the Spatio-Temporal Reflectance Prior 37
3.5.4 Nested Hierarchical Optimization 38
3.6 Experiments 38
3.6.1 Qualitative Evaluation 40
3.6.2 Quantitative Evaluation 40
3.6.3 Evaluation on 'Intrinsic Images in the Wild' Dataset 42
3.6.4 Influence of the Different Energy Terms 43
3.6.5 Runtime and Convergence 44
3.7 Applications 45
3.7.1 Dynamic Reflectance Recolouring 45
3.7.2 Editing Material Appearances 46
3.7.3 Realistic Texture Replacement 47
3.7.4 Live Video Abstraction & Stylization 48
3.8 Discussion 48
3.9 Conclusion 50

4 Live User-Guided Intrinsic Video 51
4.1 Introduction 51
4.2 Related Work 53
4.3 Overview 55
4.4 Representation 56
4.5 Interactions 57
4.5.1 Detection of Touch Points 58
4.5.2 Spatial Constraint Propagation 58
4.6 Energy 59
4.6.1 Variational Intrinsic Video Decomposition 60
4.6.2 Data-Parallel Optimization 62
4.7 Experiments 63
4.8 Interactive Applications 67
4.9 Discussion 69
4.10 Conclusion 70

5 Live Global Intrinsic Video **71**
5.1 Introduction . . . 72
5.2 Related Work . . . 73
5.3 Overview . . . 75
5.4 Method . . . 76
5.5 Base Colour Estimation . . . 77
5.5.1 Chromaticity Clustering . . . 77
5.5.2 Misclustering Correction . . . 78
5.6 Shading Decomposition . . . 79
5.6.1 Reflectance Priors . . . 80
5.6.2 Shading Priors . . . 81
5.6.3 Base Colour Refinement . . . 83
5.6.4 Handling the Sparsity-Inducing Norms . . . 84
5.7 Optimization . . . 85
5.7.1 SparseDense Splitting . . . 86
5.8 Experiments . . . 86
5.8.1 Quantitative Results . . . 87
5.8.2 Qualitative Results . . . 89
5.8.3 Comparisons . . . 94
5.8.4 Interactive Live Applications . . . 96
5.9 Discussion . . . 99
5.10 Conclusion . . . 100

II Real-time Reflectance Acquisition **101**

6 Live Intrinsic Material Estimation **103**
6.1 Introduction . . . 103
6.2 Related Work . . . 105
6.3 Overview . . . 107
6.4 Appearance Model . . . 107
6.5 Learning . . . 108
6.5.1 Synthetic ground-truth Training Corpus . . . 108
6.5.2 Physically Motivated Network Architecture . . . 109
6.5.3 Perceptual Rendering Loss . . . 112
6.5.4 End-to-End Training . . . 113
6.6 Temporal Fusion . . . 113
6.7 Experiments . . . 115

6.7.1 Qualitative Results . . . 115
6.7.2 Run-time Performance . . . 116
6.7.3 Quantitative Evaluation and Ablation Study . . . 116
6.7.4 Comparison to the state-of-the-art . . . 118
6.8 Applications . . . 119
6.9 Discussion . . . 121
6.10 Conclusion . . . 121

7 Deep Reflectance Fields **123**
7.1 Introduction . . . 123
7.2 Related Work . . . 125
7.3 Capture Setup . . . 128
7.3.1 Spherical colour gradient images . . . 129
7.3.2 Hardware and data capture . . . 129
7.4 Learning . . . 131
7.4.1 Predicting photorealistic 4D reflectance fields . . . 131
7.4.2 Training . . . 132
7.4.3 Inference . . . 134
7.5 Experiments . . . 135
7.5.1 Qualitative Comparisons . . . 135
7.5.2 Ablation study . . . 139
7.5.3 User study . . . 141
7.5.4 Environment map based relighting . . . 142
7.6 Discussion . . . 143
7.7 Conclusion . . . 143

8 Conclusion **145**
8.1 Insights . . . 145
8.2 Outlook . . . 147

References **169**

List of Figures

1.1 Photorealism in Computer Graphics 2
1.2 Personalized graphics on mobile devices 3

2.1 Environmental illumination mapping 14

3.1 Pipeline for Live Intrinsic Video 27
3.2 Chromaticity shift 30
3.3 Spatio-temporal reflectance consistency prior 31
3.4 Reflectance Clustering 32
3.5 Subdomains of the localglobal optimization approach. 36
3.6 State-of-the-art comparison on the GIRL sequence. 39
3.7 State-of-the-art comparison on the TOY sequence 39
3.8 ground-truth comparison on the SANMIGUEL sequence 41
3.9 Quantitative evaluation on the SANMIGUEL sequence 41
3.10 Influence of energy terms: ablation study 43
3.11 Influence of different priors on the SANMIGUEL sequence 43
3.12 Convergence: The residual error is always decreasing. 44
3.13 Reflectance recolouring on the GIRL sequence 45
3.14 Editing material appearances on the OBJECTS sequence 46
3.15 Photorealistic texture replacement: a virtual painting (left) is added, and a brick texture (right) applied 46
3.16 Photorealistic texture replacement: a virtual painting is added to the wall 47
3.17 Realistic texture replacement: two virtual decals are added to a box 47
3.18 Live video stylization using a cartoon-style effect. 48
3.19 Recolouring of highly textured objects 49

4.1 Piepeline for novel user-guided intrinsic video 52
4.2 User detection 58

4.3 Constant reflectance strokes improve the decomposition by moving the high-frequency shading of the cloth to the shading layer. . . . 62
4.4 Intrinsic decomposition results for a colour chart 65
4.5 Comparison to state-of-the-art intrinsic video decomposition techniques on the 'girl' dataset 65
4.6 Temporal reflectance constancy 65
4.7 The proposed approach reconstructs the reflectance of the scene. . 66
4.8 Photorealistic recolouring of a shirt. 67
4.9 Interactive object recolouring 68
4.10 Dynamic geometry-based relighting 68
4.11 The shading layer is modified to convert plaster to metal. 69

5.1 real-time decomposition of a video into reflectance, direct and indirect shading components 71
5.2 Pipeline for global intrinsic decomposition 75
5.3 Example of misclustering correction 78
5.4 Base colour refinement 83
5.5 SyntheticRoom sequence is quantitatively analysed 87
5.6 Decomposition of the ChitChat sequence 88
5.7 Decomposition of the Umbrella sequence 88
5.8 The decomposition of the Girl2 sequence 89
5.9 Decomposition of the Droid sequence 90
5.10 Live recolouring of scene surfaces in a photorealistic and globally consistent manner 90
5.11 Global intrinsic decomposition applied to two samples from the Intrinsic Images in the Wild dataset of Bell *et al.* (2014). 91
5.12 Results on the Box sequence, with and without the novel sparsity-based misclustering correction 91
5.13 Comparison of the proposed global intrinsic decomposition to the state-of-the-art techniques on the Toy sequence 92
5.14 Evaluation of the soft-colour-Retinex weight of the monochromatic shading term on the ChitChat sequence 92
5.15 Comparison of the global intrinsic decomposition result on the Droid sequence, with and without the shading sparsity prior 93
5.16 The global intrinsic decomposition is qualitatively compared to the ground-truth on the synthetic Cornell sequence 93
5.17 Comparison to the state-of-the-art on the Paper sequence 94

5.18 Decomposition result on the SyntheticRoom sequence 95
5.19 Comparison of recolouring results to state-of-the-art technique on the Kermit sequence 96
5.20 Recolouring result on the Cup sequence 96
5.21 Comparison to commercial recolouring softwares 97
5.22 Comparison to state-of-the-art techniques on the Cart sequence . 98

6.1 Pipeline for real-time estimation of material parameters from a single monocular colour image 105
6.2 Synthetic ground-truth training corpus 109
6.3 Network architecture of the sub-networks of the proposed technique. 110
6.4 Real world material estimation results based on a single colour image 114
6.5 Material estimates on the dataset of Rematas *et al.* (2016). 114
6.6 Specular decomposition layer from a single colour image 115
6.7 Confusion matrix of shininess prediction for classification (left) and regression of log-shininess (right). 117
6.8 Material shininess classification bins 117
6.9 Material estimation and transfer comparison 119
6.10 Comparison to the state-of-the-art approach 119
6.11 Cloning of real world materials on virtual objects in an illumination-consistent fashion 120

7.1 The Lightstage setup 128
7.2 An environment map can be approximated with the 331 lighting directions of the Light Stage 129
7.3 Network architecture 131
7.4 Effect of different training losses on the final results 132
7.5 Alignment Loss 133
7.6 Examples of gradient input images, inferred OLAT images and ground truth 136
7.7 Generalization w.r.t. light direction 137
7.8 Generalization w.r.t. viewpoint 138
7.9 Comparison of regressed OLAT images with different methods . . . 139
7.10 Relighting Comparisons – Comparisons with other state of art approaches. 139
7.11 Dynamic results 140
7.12 Comparisons with different input modalities 141
7.13 Relighting results with HDRI lighting environments 142

LIST OF FIGURES

List of Tables

3.1 Quantitative comparison on the SANMIGUEL sequence: The obtained decompositions have a lower error (bold) than previous work. . . . 42
3.2 Runtime performance of Live Intrinsic Video 44

4.1 Run time of the proposed user-guided intrinsic video 64

5.1 user-interactions required for all sequences. Note that most sequences do not require any user-interaction (bottom half of the table). . . . 91

6.1 Quantitative evaluation on a test set of 4,990 synthetic images. The column Shininess Exponent shows the accuracy of exponent classification, reported as percentage classified in the correct bin and the adjacent bins. The last three columns show the direct parameter estimation mean square error over the full test set. Please note that the error on shininess is evaluated in log-space to compensate for the exponential bias. 116

7.1 Quantitative evaluations on test sequences of subjects. Photometric error is measured via the ℓ_1-norm. Keeping architecture fixed, the proposed loss function is compared with the other baselines. Significantly lower MSE with the ground-truth is obtained while the SSIM score is similar to the other networks. Do note that these statistical measured often do not quantify well the subjective photorealism of the images. 140

LIST OF TABLES

Chapter 1

Introduction

1.1 Motivation

Creating simulated models of real world objects and scenes has been a cherished desire for a large part of the human history. Paintings, sculptures or graphical media – such replicas of the real world have enabled us to communicate information and ideas over large distances and understand them better through improved visualization. Good visual models have benefited teachers, storytellers, scientists and traders alike. Thus, such simulated models have been the cornerstones of human civilization and knowledge economy.

The digital revolution, particularly the advent of computer graphics, has given a large impetus to this phenomenon. Now, we can not only replicate reality, but also *create* new realities or *manipulate* them in desirable ways. From movies to video games, mechanical design to brand endorsements, food advertisements to furniture catalogues, computer graphics has changed the way we share and visualize stories and ideas. Photorealism in graphics, defined as a quality of a model to be indistinguishable from its real version, has always been challenging to researchers, particularly because it involves deceiving one the most finely tuned of the human senses – the visual perception system. Human vision is wonderfully adept at combining information from various visual cues and gaining a comprehensive understanding of the world around it. However, this makes the job of anyone trying to simulate the appearance of the real world that much more difficult.

Attempts at mimicking the appearance of the real world with digital models have challenged our knowledge of the *physical* nature of visual appearance. While optical physics has pushed our understanding of how light interacts with physical media, the study of geometric vision has helped us to decipher the rules that govern how this

1.1 Motivation

Figure 1.1: *Left: High-quality rendering of a Tibetan monk character by 3D artist Gui Wenlong. (http://www.artstation.com/artwork/dOy4GJ), Right: 3D compositing before/after full rendering in a scene from the movie 'Jungle Book', (copyright Disney Corporation).* Photorealistic computer graphics has significantly enhanced large-scale communication of ideas by enabling artists to create new realities. This quality of realism is achieved through large scale projects requiring skilled manual artwork.

reflected light influences the images that we see. However, it is the computational graphical models of these processes that have helped us achieve the ability to create new photorealistic digital scenes.

Starting with the rendering equation of Kajiya (1986), numerous advancements have been made that simulate the light transport image-formation model. The rendering equation identifies three basic scene components that influence image formation – geometry, reflectance and illumination. Naturally, each of these components has received tremendous attention from graphics researchers over the years, resulting in accurate and efficient representations, equipment for acquisition and algorithms for manipulation. An algorithm that integrates these scene components based on the light transport model to generate 2D images is termed as *rendering* and the inverse process of estimating these scene components from 2D images is known as *inverse rendering*.

Inverse rendering has traditionally required complex optical equipment and tedious manual effort, but has been rewarding as it has led to photorealistic digital renderings of real world specimen (Figure 1.1). But the growing parallel computing capacities of modern computers and increasing capabilities of modern miniaturized

Figure 1.2: *Left: Ikea's Place mobile app, Center-Bottom: Porsche's Mission E mobile app, Center-Top: Microsoft's Hololens AR headset, Right: Microsoft's Holoportation project.* In the last decade, high-quality computer graphics has transcended large screens to a more personalized format on mobile phones and head-mounted displays. While this is primarily possible due to improved geometric understanding of the environment, the photorealism of the graphics is limited by a unavailability or poor quality of the scene *reflectance* and *illumination*.

sensors and electronics has brought in a new change to the world of computer graphics in the last decade in the form of 'personalized graphics'. Photorealistic graphics have come down from billboards and television screens to mobile devices. Personal augmented/mixed reality (AR/MR) devices have become ubiquitous. These devices generally consist of a 'see-through' display either as a head or eye mounted device with a projection mechanism or a simple hand held display like a mobile camera screen, as seen in Figure 1.2. They seamlessly superimpose graphical content into the real world as seen through the display. They bring with them the promise of a tectonic shift in media and communication technology - an entirely new way of interacting with the world. By being able to process scene information in real-time, they can acquire, interpret and modify the scene to achieve novel means of communication and visualization.

This rise of mixed reality technology is driven by greater environmental awareness afforded to these devices by advances in several robust real-time tracking and mapping algorithms from few or a single camera, such as – SLAM (simulataneous localization and mapping) (Saputra *et al.*, 2018), pose estimation (Gong *et al.*, 2016; Marchand *et al.*, 2016) and geometry reconstruction (Zollhöfer *et al.*, 2018). Most AR devices today rely on a version of these algorithms.

Interestingly, all of these algorithms are aimed at acquiring information about the scene geometry. While the theory and algorithms that drive geometry acquisition in real-time have progressed rapidly, the same cannot be said about the other fundamental scene components from the rendering equation – reflectance and illumination. There exists very little scientific works prior to the beginning of this

thesis that address the problem of estimating these components in real-time.

This is the first comprehensive scientific thesis that deals with acquiring scene reflectance and illumination in real-time in a live causal setting from a single or very few cameras, in order to enable photorealism in real-time computer graphics applications on mixed reality devices. The thesis introduces the design and implementation of novel representations, algorithms and systems for reflectance and illumination acquisition. Particularly, the thesis focuses on generalizing the algorithms to work on unstructured scenes and in-the-wild settings. The effectiveness of the proposed solutions is extensively established through qualitative and quantitative experimentation and by demonstrating several real world augmented reality applications.

1.2 Overview

The primary goal of this thesis is to explore the solution space of inverse rendering problems in real-time and live scenarios, with few or even a single camera. This is particularly challenging on the following counts:

1. Finding efficient representations for reflectance and illumination.
2. The high-dimensional and non-linear nature of the resulting problem.
3. Highly under-constrained nature of the equations due to small number of measurements.

This thesis explores a range of representations and algorithms. It begins with a simple linear approximation of the rendering equation under the single-bounce Lambertian reflectance assumption. In this case the problem collapses to the classical intrinsic scene decomposition problem which is solved for the very first time from a monocular video steam in unconstrained settings at real-time frame rates. This is achieved by designing fast and robust spatio-temporal priors and solving the resulting sparse non-linear optimization problem using a novel iterative GPU optimizer.

This line of work is futher extended to 3D by explicitly incorporating geometry reconstruction of static scenes into the algorithm. This additionally allows for better novel-view constraints and user-guidance directly on the scene. The interactive user-guidance cues are automatically propagated across the scene using the reconstructed geometry, to further enhance the quality of intrinsic decomposition.

Next, it deals with the more challenging multi-bounce light reflection case by solving for the global illumination in a scene using novel sparsity priors. This leads

to a combination of dense and sparse equations in reflectance and illumination and is solved using a custom-made dense-sparse alternating GPU optimizer.

Going forward, the Lambertian assumption is dropped to deal with more general materials. For the first time, a solution for regressing the reflectance of an object of general specularity from a single image in real-time is demonstrated. This is achieved using convolutional neural networks driven by a novel reconstruction loss that estimates consistent material properties and intrinsic images for an object under unconstrained lighting using the Blinn-Phong model.

Finally, the complex light transport effects exhibited by the human face are explored with a data-driven reflectance basis.The first real-time technique to regress a full image-space reflectance field of dynamic human face performances in a Light Stage is presented. This is achieved using a convolutional neural network with a novel task-specific loss designed to capture specularities and light-direction and view-direction dependent effects.

Several real world applications are exhibited that push the state-of-the-art of inverse rendering.

1.3 Structure

This thesis is divided into eight chapters:

- Chapter 1 motivates the topic of this thesis, provides an overview of the work, outlines the structure of exposition and stresses the main technical contributions.

- Chapter 2 describes the fundamental concepts of inverse rendering and the mathematical notation that is used throughout this thesis.

- Chapters 3 to 7 present the main technical contributions. The related past literature is discussed at the beginning of each chapter, extensive experimentation is reported and various applications of the proposed work are demonstrated at the end.

- Chapter 8 summarizes the core contributions and provides an outlook for future work.

1.4 Contributions

This section summarizes the main contributions of this dissertation.

The main contributions of Chapter 3 (published as Meka *et al.* (2016)) are:

- The first real-time algorithm to decompose a live monocular video stream into high-quality reflectance and shading layers.
- A novel formulation for the intrinsic video decomposition problem that combines local spatial and global spatio-temporal priors tailored to produce high-quality and temporally consistent video decompositions in real-time.
- A new data-parallel solver for mixed l_2–l_P optimization problems based on iteratively reweighted least squares (IRLS).

The main contributions of Chapter 4 (published as Meka *et al.* (2017b)) are:

- The first real-time algorithm to acquire reflectance fused 3D geometry of static scenes.
- A reflectance-fused volumetric scene representation that uses user-provided 3D strokes for refining the intrinsic decomposition.
- A novel-view projection based temporal constraint for improved temporal stability of intrinsic video decomposition.

The main contributions of Chapter 5 (published as Meka *et al.* (2019b)) are:

- The first real-time algorithm to decompose a live monocular video stream into reflectance and direct and indirect illumination components.
- A sparsity-based automatic estimation of the underlying reflectance when a user identifies regions of strong interreflections.
- A novel parallelized sparse-dense optimizer to solve a mixture of high-dimensional sparse problems jointly with low-dimensional dense problems at real-time frame rates.

The main contributions of Chapter 6 (published as Meka *et al.* (2018)) are:

- The first real-time algorithm to estimate isotropic material reflectance properties of an object surface from a single RGB camera.

- A novel learning based pipeline along with a reconstruction loss based on consistent material reflectance regression and intrinsic image estimation of an object image.
- Real-time high-frequency illumination estimation of the scene by looking at a single object using a depth sensor.

The main contributions of Chapter 7 (published as Meka *et al.* (2019a)) are:

- The first real-time algorithm to acquire the full reflectance field of a human facial performance in a Light Stage.
- A learning-based formulation that maps two spherical gradient images to the full 'one-light-at-a-time' (OLAT) reflectance basis.
- A task-specific perceptual loss trained to pick up specularities and high-frequency details and an alignment loss that robustly handles the small misalignments between the training input-output pairs.

1.4 Contributions

Chapter 2

Technical Background

2.1 Image Formation

Image formation can be described as the process of rays of light emitted by a light source, reflected by objects in a scene and finally recorded by a sensor. While the description sounds simple, there are several underlying phenomena such as emission, transmission, reflection, refraction, scattering and digitization that contribute significantly to the final image, each of which has led to a separate field of study.

Optical physics has made great strides in understanding the propagation and interaction of light in different media. Radiometry has excelled in designing devices and techniques to measure light characteristics. Electronics has developed processes and hardware to faithfully convert scene radiance to machine-interpretable digital signals. Computer graphics has specialized in designing efficient computational models to simulate the processes involved in light transport and computer vision has extensively studied the statistics and patterns inherent in 2D images.

Each of these fields has borrowed from and built on top of each other. The scope of this thesis is confined to addressing the image deconstruction problem known in the field computer vision using computational models from graphics, while also very briefly touching on associated theories and machinery from the other allied fields.

2.1.1 Rendering Equation

In computer graphics, image formation is primarily dealt with by modelling the parameters of light reflection by physical surfaces, while assuming that other factors are known and fixed. A comprehensive graphical model for reflection was introduced simultaneously by Immel *et al.* (1986) and Kajiya (1986) in the form of the *rendering*

equation.

$$\begin{aligned}\mathbf{L}_o(\mathbf{x},\boldsymbol{\omega}_o) &= \mathbf{L}_e(\mathbf{x},\boldsymbol{\omega}_o) + \mathbf{L}_r(\mathbf{x},\boldsymbol{\omega}_o) \\ &= \mathbf{L}_e(\mathbf{x},\boldsymbol{\omega}_o) + \int_{\boldsymbol{\Omega}} \mathbf{f}(\mathbf{x},\boldsymbol{\omega}_i,\boldsymbol{\omega}_o)\mathbf{L}_i(\mathbf{x},\boldsymbol{\omega}_i)(\boldsymbol{\omega}_i \cdot \mathbf{n})\, d\boldsymbol{\omega}_i. \end{aligned} \tag{2.1}$$

The rendering equation expresses the radiance $\mathbf{L}_o \in \mathbb{R}^3$ leaving a surface point $\mathbf{x} \in \mathbb{R}^3$ (with normal $\mathbf{n} \in \mathbb{S}^2$) in direction $\boldsymbol{\omega}_o \in \mathbb{S}^2$ as the sum of emitted $\mathbf{L}_e \in \mathbb{R}^3$ and reflected radiance $\mathbf{L}_r \in \mathbb{R}^3$. The reflected radiance $\mathbf{L}_r$ is a function of the illumination $\mathbf{L}_i \in \mathbb{R}^3$ over the hemisphere $\boldsymbol{\Omega}$ of incoming directions $\boldsymbol{\omega}_i \in \mathbb{S}^2$ and the material's BRDF $\mathbf{f} : \mathbb{R}^3 \times \mathbb{S}^2 \times \mathbb{S}^2 \to \mathbb{R}$ at point $\mathbf{x}$ with a local normal $\mathbf{n}$.

Based on the Equation (2.1), the scene radiance and hence the recorded image is influenced by three independent physical parameters:

- *surface reflectance* ($\mathbf{f}$),
- *illumination* ($\mathbf{L}_i$)
- *scene geometry* ($\mathbf{x}$, $\mathbf{n}$)

Given these parameters, the exact irradiance at an image pixel can be computed by following paths of rays emitted from light sources and reflected by scene surfaces before ending up at the pixel location, in a process known as ***rendering***. This involves solving the rendering equation at every point of intersection of a light ray with a physical surface.

2.1.2 Inverse Rendering

The rendering equation allows us to generate photorealistic images of 3D graphical models with predefined geometry, reflectance and illumination. The same equation can also be used to invert this process in real world images to compute the three components from a 2D image. This process is known as ***inverse rendering***. But inverse rendering is a severely under-constrained problem. Given only a single pixel value, it is to be factorized into three different components. Such an equation will have infinite solutions. The problem is made even more intractable by the fact the two of the components – reflectance and shading – are continuous high-dimensional fields. While geometry can be discretized and represented at a given 3D point using representations such as meshes, voxel grids and point clouds, finding low-dimensional representations for reflectance and illumination is a challenge. Thus, it is beneficial to look at these problems separately.

Estimation of 3D geometry from images, under simple assumptions on reflectance and illumination, has been studied extensively over the years with tremendous progress over the last decade (Zollhöfer *et al.*, 2018). Several methods have been developed that use single, stereo, multi-view and depth images using cues such as shading, motion, disparity, defocus or a combination of them to perform 3D reconstruction from 2D data (Hartley & Zisserman, 2003; Moons *et al.*, 2010). More recently, machine learning methods have also been used to learn to solve 3D reconstruction by looking at a vast amount of 2D data (Han *et al.*, 2019). While geometry is an important problem to solve, the solution to geometry reconstruction falls outside the scope of this thesis.

The term *inverse rendering* is more generally used to refer to the problem of estimating reflectance and illumination from 2D images. Reconstructing these components from 2D data is a challenging and ill-posed problem (Patow & Pueyo, 2003; Ramamoorthi & Hanrahan, 2001c; Yu *et al.*, 1999). Most approaches need to make strong assumptions, such as the availability of an RGBD camera (e.g. Guo *et al.*, 2017; Wu *et al.*, 2016), strong priors such as a data-driven BRDF model (Lombardi & Nishino, 2016b) or flash lighting (Li *et al.*, 2018b; Nam *et al.*, 2018), knowledge of geometry (Azinovi *et al.*, 2019; Dong *et al.*, 2014; Li *et al.*, 2017; Marschner & Greenberg, 1997) or a specific object class (Georgoulis *et al.*, 2017b; Liu *et al.*, 2017). This thesis will deal with solving these problems with a small number or a single sensor, in real-time settings, while generalizing to unconstrained scenarios. Before delving into solutions to these problems in the next chapters, the problem of finding representations for reflectance and illumination are discussed in more detail in the next sections.

2.2 Reflectance

The reflectance function (also called bi-directional reflectance distribution function (BRDF)) $\mathbf{f}(\boldsymbol{\omega}_i,\boldsymbol{\omega}_o)$ is an intrinsic property of a material that defines the behaviour of light reflection at the material's surface. The function takes the incoming light direction $\boldsymbol{\omega}_i$ and outgoing light direction $\boldsymbol{\omega}_o$ as input and outputs the ratio of the radiance to the irradiance. Since each direction $\boldsymbol{\omega}$ has two degrees of freedom parameterized by azimuth angle ϕ and zenith angle θ, the BRDF as a whole is a 4D function. In case of a non-uniform material surface, the 4D BRDF function itself changes over the surface geometry as $\mathbf{f}(\boldsymbol{\omega}_i,\boldsymbol{\omega}_o,\mathbf{x})$ and hence is a 6D function.

2.2.1 Reflectance models

The space of possible reflectance functions is very large. Several simplifying models have been proposed in computer graphics to efficiently simulate reflectances of real world objects. While these models may only be approximations of real world BRDFs, their low-dimensional, explicit form and differentiability makes them more tractable to solve the highly under-constrained inverse rendering problem. In the following sections, the BRDF models that are relevant to this thesis are discussed.

2.2.1.1 Lambertian model

Lambertian reflectance defines an ideal diffusely reflecting surface, i.e., the function has a constant value with respect to the incoming ($\boldsymbol{\omega}_i$) and outgoing ($\boldsymbol{\omega}_o$) light directions. In this case, the rendering equation boils down to

$$\mathbf{L}_o(\mathbf{x},\boldsymbol{\omega}_o) = \mathbf{f}(\mathbf{x},\boldsymbol{\omega}_i,\boldsymbol{\omega}_o) \int_{\Omega} \mathbf{L}_i(\mathbf{x},\boldsymbol{\omega}_i)(\boldsymbol{\omega}_i \cdot \mathbf{n})\, d\boldsymbol{\omega}_i \tag{2.2}$$

$$= \mathbf{m}_d(\mathbf{x}) \int_{\Omega} \mathbf{L}_i(\mathbf{x},\boldsymbol{\omega}_i)(\boldsymbol{\omega}_i \cdot \mathbf{n})\, d\boldsymbol{\omega}_i \tag{2.3}$$

where $\mathbf{f}(\mathbf{x},\boldsymbol{\omega}_i,\boldsymbol{\omega}_o)$ is reduced to a constant in the light directions and is referred to as *diffuse albedo* ($\mathbf{m}_d$).

2.2.1.2 Phong model

The Phong model (Phong, 1975) is an empirical model for reflectance.

$$\mathbf{f}_{phong}(\mathbf{x},\mathbf{n},\boldsymbol{\omega}_i,\boldsymbol{\omega}_o) = \underbrace{\mathbf{m}_d(\mathbf{x})}_{\text{diffuse}} + \underbrace{\mathbf{m}_s(\mathbf{x})(\boldsymbol{\omega}_o \cdot \mathbf{r})^s}_{\text{specular}}. \tag{2.4}$$

Here, $\mathbf{m}_d \in \mathbb{R}^3$ is the diffuse albedo and $\mathbf{m}_s \in \mathbb{R}^3$ the specular albedo. The direction vector $\mathbf{r}$ is calculated as the reflection of $\boldsymbol{\omega}_i$ on the surface characterized by the surface normal $\mathbf{n}$ using $\mathbf{r} = 2(\boldsymbol{\omega}_i \cdot n)n - \boldsymbol{\omega}_i$.

The Phong model consists of a diffuse albedo similar to the Lambertian model, but additionally has a 'view-dependent' specular component. The specular component consists of a specular albedo times a lobe with a cosine fall-off, centred around the reflection direction $\mathbf{r}$. The scalar exponent $s \in \mathbb{R}$ determines the width of the

specular lobe, and thus the 'shininess' of the material. Since the fall-off of the lobe is uniform in all directions, the model is said to be 'isotropic'.

2.2.1.3 Blinn-Phong model

The Blinn-Phong model (Blinn, 1977) is also called the modified Phong model. It is a more computationally efficient version of the Phong model.

$$\mathbf{f}_{blinn-phong}(\mathbf{x},\mathbf{n},\boldsymbol{\omega}_i,\boldsymbol{\omega}_o) = \underbrace{\mathbf{m}_d(\mathbf{x})}_{\text{diffuse}} + \underbrace{\mathbf{m}_s(\mathbf{x})(\mathbf{h}\cdot\mathbf{n})^s}_{\text{specular}}. \tag{2.5}$$

The halfway vector $\mathbf{h} = \frac{\boldsymbol{\omega}_i+\boldsymbol{\omega}_o}{\|\boldsymbol{\omega}_i+\boldsymbol{\omega}_o\|}$ depends on the light direction $\boldsymbol{\omega}_i$ and the viewing direction $\boldsymbol{\omega}_o$. The other variables carry the same meaning as the Phong model. The Blinn-Phong model has been empirically shown to reproduce more accurate reflectance effects than the Phong model for many types of surfaces (Ngan *et al.*, 2004).

Both the Phong and the Blinn-Phong models provide for a linear decomposition of an image into diffuse and specular components, which can further be expressed as a product of an albedo parameter and a shading image. Such a decomposition of the Blinn-Phong model is exploited in Chapter 6 to design an effective strategy to measure the parametric reflectance of general objects.

2.3 Illumination

In computer graphics, illumination is represented using a large class of models, from a simple directional light source whose contribution to a (diffuse) surface point can be computed with a dot product, to a complex light emitting near-scene object which requires complex ray-tracing, radiosity and Monte Carlo computations to simulate.

In this thesis, illumination in a scene is always assumed to be from an infinite distance. Such an assumption is realisitic in cases where light sources are significantly farther away from the scene than the relative distances of points within the scene. This is usually the case in the experiments described in this thesis.

Apart from a simple point or directional light source, illumination in the scene has been represented either by an environment map or spherical harmonics representation. These are described below.

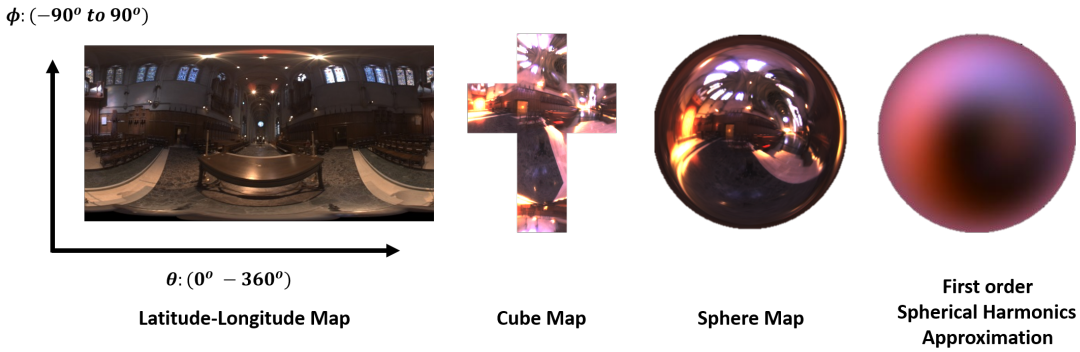


Figure 2.1: The 'Grace Cathedral' environment map (courtesy of Debevec (1998)) expressed in various 2D representations. A sphere map corresponding to the low-frequency spherical harmonics approximation is also shown.

2.3.1 Environment map representation

An environment map is a function that maps an input direction $\boldsymbol{\omega}_i$ to a coloured light intensity. This mapping is stored in the form of a 2D 'latitude-longitude' rectangular image, as shown in Figure 2.1. The azimuth angle θ through $0°$ to $360°$ is linearly mapped to the horizontal axis (latitude) of the image and the zenith angle ϕ through $-90°$ to $90°$ is mapped to the vertical axis (longitude). Such a representation is used in Part II of the thesis to represent the lighting in the scene. Alternatively, a cube-map or a sphere-map could also be used. Note that this mapping has an explicit form and is a high-dimensional representation with million (of pixels) of free variables, hence not a very tractable representation for regression tasks.

2.3.2 Spherical harmonics illumination

Spherical harmonics is a frequency-space basis for representing functions defined over the unit sphere. They are a spherical analogue of the 1D Fourier series. Since distant illumination such as the environment map is defined over a unit sphere, spherical harmonics have been very successfully used for approximating environmental illumination (Ramamoorthi & Hanrahan, 2001a).

If the incident direction $\boldsymbol{\omega}_i$ is represented using the standard spherical parameterization,

$$\boldsymbol{\omega}_i = (\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta) \tag{2.6}$$

the real spherical harmonic basis functions are defined as:

$$y_l^m(\theta,\phi) = \left\{ \begin{array}{ll} \sqrt{2}K_l^m \cos m\phi P_l^m(\cos\theta) & \text{if } m>0, \\ K_l^0 P_l^0(\cos\theta) & \text{if } m=0, \\ \sqrt{2}K_l^m \sin -m\phi P_l^{-m}(\cos\theta) & \text{if } m<0. \end{array} \right\} \tag{2.7}$$

where K_l^m are the normalization constants:

$$K_l^m = \sqrt{\frac{(2l+1)(l-|m|)!}{4\pi(l+|m|)!}} \tag{2.8}$$

and P_l^m are the associated Legendre polynomials. The basis functions are indexed according to two integer constants, the order, l, and the degree, m[2]. These satisfy the constraint that $l \in N$ and $-l \leq m \leq l$; thus, there are $2l+1$ basis functions of order l. The order l determines the frequency of the basis functions over the sphere.

The first order spherical harmonics are given as:

$$\begin{array}{ll} l=0 & \left\{ \begin{array}{ll} y_0^0(\theta,\phi) & = \sqrt{\frac{1}{4\pi}} \end{array} \right\} \\ l=1 & \left\{ \begin{array}{ll} y_1^{-1}(\theta,\phi) = & \sqrt{\frac{3}{4\pi}\sin\phi\sin\theta} \\ y_1^0(\theta,\phi) = & \sqrt{\frac{3}{4\pi}\cos\theta} \\ y_1^1(\theta,\phi) = & \sqrt{\frac{3}{4\pi}\cos\phi\sin\theta} \end{array} \right\} \\ l=2 & \left\{ \begin{array}{ll} y_2^{-2}(\theta,\phi) = & \sqrt{\frac{15}{4\pi}\sin\phi\cos\theta\sin^2\theta} \\ y_2^{-1}(\theta,\phi) = & \sqrt{\frac{15}{4\pi}\sin\phi\sin\theta\cos\theta} \\ y_2^0(\theta,\phi) = & \sqrt{\frac{5}{16\pi}(3\cos^2\theta-1)} \\ y_2^1(\theta,\phi) = & \sqrt{\frac{15}{4\pi}\cos\phi\sin\theta\cos\theta} \\ y_2^2(\theta,\phi) = & \sqrt{\frac{15}{16\pi}(\cos^2\phi-\sin^2\phi)\sin^2\theta} \end{array} \right\} \end{array}$$

The incident illumination $\mathbf{L}_i$ can be expressed as a linear combination of the spherical harmonics basis functions as:

$$\mathbf{L}_i(\boldsymbol{\omega}_i) = \sum_{l=0}^{\inf} \sum_{m=-l}^{l} y_l^m(\theta,\phi)\mathbf{c}_l^m, \tag{2.9}$$

where the co-efficients $\mathbf{c}_l^m$ can be computed by projecting $\mathbf{L}_i$ onto each basis function y_l^m

$$\mathbf{c}_l^m = \int_{\mathbf{\Omega}_{4\pi}} y_l^m(\theta,\phi)\mathbf{L}_i(\boldsymbol{\omega}_i)d\omega_i. \tag{2.10}$$

Ramamoorthi & Hanrahan (2001a) have shown that using only the nine first order spherical harmonic co-efficients to approximate an environment map, efficient rendering of diffuse objects can be performed with less than 1% average error. Conversely, Ramamoorthi & Hanrahan (2001b) estimate the first order spherical harmonic co-efficients of the incident illumination from an image of an object with Lambertian reflectance. Similarly, in Chapter 6, a low frequency representation of the environment map is reconstructed from a diffuse shading image of general class of objects by computing the first order spherical harmonics co-efficients.

2.4 Intrinsic Decomposition

Intrinsic image decomposition, first proposed by Barrow & Tenenbaum (1978a), is a simplified version of the inverse rendering problem. Consider the Lambertian reflectance model previously described in Equation (2.2). While the reflectance is reduced to a constant value diffuse albedo ($\mathbf{m}_d$), the rest of the integral is also constant in the outgoing direction $\boldsymbol{\omega}_o$:

$$\mathbf{L}_o(\mathbf{x},\boldsymbol{\omega}_o) = \mathbf{m}_d(\mathbf{x}) \int_{\mathbf{\Omega}} \mathbf{L}_i(\mathbf{x},\boldsymbol{\omega}_i)(\boldsymbol{\omega}_i \cdot \mathbf{n})\, d\boldsymbol{\omega}_i \tag{2.11}$$

$$= \mathbf{m}_d(\mathbf{x})\, \mathbf{s}(\mathbf{x},\boldsymbol{\omega}_i) \tag{2.12}$$

where $\mathbf{x}$ is the pixel location and '$\mathbf{s}$' is referred to as shading. The shading image ($\mathbf{s}$) encompasses the integral over the hemisphere Ω. Thus, it is influenced by both the local geometry $\mathbf{n}$ and the incident illumination $\mathbf{L}_i$. For a single image where the illumination is static, the intrinsic decomposition is then defined as:

$$\mathbf{L}_o(\mathbf{x}) = \mathbf{m}_d(\mathbf{x})\, \mathbf{s}(\mathbf{x}) \tag{2.13}$$

Hence, intrinsic decomposition is the problem of factorization of every pixel in an image into a diffuse reflectance and a shading component. While still under-

constrained, this simple formulation allows for various image editing applications, as will be shown in the Part I of this thesis.

Part I

Real-time Intrinsic Decomposition

Chapter 3

Live Intrinsic Video

One of the simplest yet most widely applicable formulations of the inverse rendering problem is that of intrinsic scene decomposition. As described in the previous chapter Section 2.1, intrinsic decomposition, under a diffuse reflectance assumption, aims to find a per-pixel factorization of a given image into reflectance and shading layers. Although the problem has received a great deal of attention from computer graphics researchers in the past, its applicability has been limited due to the computational complexity and resulting large processing times of the state-of-the-art approaches. This chapter presents the first technique that not only solves the problem in real-time, but also finds novel solutions to the challenges of causality and spatio-temporal consistency that arise from performing the decomposition for live video streams (Meka *et al.*, 2016). Several new photorealistic augmented reality and video editing applications enabled by this approach are demonstrated.

3.1 Introduction

Separating a video stream into its reflectance and shading layers is a fundamentally ambiguous and challenging inverse problem, but a solution has many potential applications. The availability of such a decomposition is for example the basis of a large variety of video editing tasks like realistic recolouring, relighting and texture editing. Having a fast real-time solution to this fundamental problem has big ramifications especially in the context of augmented reality since this allows to apply such modifications, in particular photorealistic texture and appearance editing, directly to live video footage.

Consider the simpler problem of computing the decomposition of a single input image. Given an image $\mathbf{I}$ (or single frame of a video), under a Lambertian scene

reflectance assumption, a decomposition is defined at every pixel $\mathbf{x}$, such that the product of reflectance $\mathbf{R}(\mathbf{x}) \in \mathbb{R}^3$ and shading $S(\mathbf{x}) \in \mathbb{R}$ is equal to the corresponding input observation:

$$\mathbf{I}(\mathbf{x}) = \mathbf{R}(\mathbf{x}) \times S(\mathbf{x}). \tag{3.1}$$

Note that the shading is modelled using the scalar quantity $S(\mathbf{x})$, based on the assumption of a white illuminant, as in previous work. Recovering the reflectance and shading image from such input constraints is ill-posed, since this problem is severely under-constrained. Equation (3.1) only provides three constraints for the four unknowns that define the reflectance $\mathbf{R}(\mathbf{x})$ and shading $S(\mathbf{x})$. This fundamental ambiguity is an inherent property of all *intrinsic decomposition* problems. State-of-the-art approaches tackle this problem by incorporating sophisticated local spatial priors that constrain the solution to a suitable subspace. These priors are based on assumptions about the typical variations encountered in reflectance and shading images. A lot of approaches Gehler *et al.* (2011); Horn (1974); Tappen *et al.* (2005) exploit the smoothness and sparsity that is often encountered in shading and reflectance images, respectively. The reflectance sparsity assumption is especially valid for most man-made objects and scenes, since these are normally composed of a small number of materials, but both assumptions might fail if more complex natural scenes are encountered.

Decompositions of such complex natural scenes can still be obtained based on more powerful discriminative priors learned from collections of training data (Barron & Malik, 2015a; Zhou *et al.*, 2015). While these approaches handle natural scenes well, they do not easily generalize to types of scenes not contained in the training data. Similarly, multi-view decomposition approaches cope with the complexity of natural scenes by exploiting multiple views of the same scene (Duchêne *et al.*, 2015; Laffont *et al.*, 2013), but these are not always available, and difficult to capture for video.

Lee *et al.* (2012) and Chen & Koltun (2013) proposed approaches that exploit simultaneously captured depth cues to resolve the ambiguities in the intrinsic decomposition problem. While their results are promising, depth information is often not easily available, especially for legacy video footage or for a live stream captured by a webcam that has to be processed at real-time frame rates.

Current state-of-the-art approaches for the intrinsic image (Barron & Malik, 2015a; Bell *et al.*, 2014; Gehler *et al.*, 2011; Li & Brown, 2014; Shen *et al.*, 2011; Zhao *et al.*, 2012) or video decomposition (Bonneel *et al.*, 2014; Kong *et al.*, 2014; Ye *et al.*, 2014) problem have prohibitively high runtimes of several minutes to hours per

frame. This makes the scene-specific parameters of these approaches hard to tune given their slow computation times. Additionally, these approaches are restricted to slow offline scenarios, where pre-recorded data is available in advance. Therefore, it is not possible to apply these techniques in the context of live applications, such as augmented reality, that require real-time processing.

Bonneel *et al.* (2014) proposed the first interactive technique that decomposes a video frame in half a second. This technique is unsuitable for the decomposition of live video streams, since it requires a slow offline pre-processing step to calculate the optical flow of the sequence. Yet, for pre-recorded data, this method offers a significant speed-up compared to previous methods. This impressive improvement in speed now allows for interactive parameter tuning, but still falls one order of magnitude short of the performance required for real-time augmented reality applications. In addition, the method relies on user-provided input in the form of scribbles, which are infeasible to provide in a real-time context.

Here, the first approach for real-time intrinsic video decomposition is proposed. The approach obtains temporally coherent decompositions at real-time frame rates without the need for explicit correspondence search. The resulting variational optimization problem is tackled using a specifically tailored data-parallel optimization strategy. High-quality decompositions are obtained even for challenging real world video sequences at the capturing rate of the input device, without requiring any user input. The main contributions are as follows:

- The first real-time algorithm to decompose live video streams into high-quality reflectance and shading layers.
- A novel formulation for the intrinsic video decomposition problem that combines local spatial and global spatio-temporal priors tailored to produce high-quality and temporally consistent video decompositions in real-time.
- A new data-parallel solver for mixed $\ell_2\ell_p$-optimization problems based on iteratively reweighted least squares (IRLS).

The approach does not require user scribbles, unlike many state-of-the-art offline approaches, yet it achieves comparable and even better results. The possibilities opened up by live intrinsic video decomposition are demonstrated by several live video editing applications, including material editing, recolouring, retexturing and stylization.

3.2 Related Work

The discussion of related work here is constrained to intrinsic decomposition methods that compute reflectance and shading layers. Many intrinsic image decomposition techniques were proposed in the past, but only very few video techniques exist that master the additional difficulty of ensuring temporally coherent results. This approach is the first to run at real-time frame rates.

Retinex and Local Priors Land & McCann (1971) suggested the Retinex approach that locally classifies edges of a grayscale image into shading or reflectance edges based on the assumption that stronger edges correspond to reflectance and weaker to shading variation. Many variants of similar and derived local edge cues have since been used (Jiang *et al.*, 2010), for instance with learned edge classifiers (Bell & Freeman, 2001; Tappen *et al.*, 2005). Retinex assumptions are also often part of more complex non-local methods. Bonneel *et al.* (2014) decompose edges into their contributing reflectance and shading components instead of simply labeling them. They use local chromaticity cues to guide the separation, and enforce sparsity on reflectance edges and smoothness on illumination edges using a hybrid $\ell_2\ell_p$-optimization strategy. This method uses similar local terms, but performs the decomposition directly on image colours instead of gradients, which avoids the integration of the gradient-domain reflectance and shading images. More recently, Bi *et al.* (2015) use a similar energy, with local colour differences in *Lab*-space used to inversely weigh the local sparsity term for reflectance estimation. Methods based only on such local cues produce decent results on simple scenes with a single segmented object, as shown in survey of Grosse *et al.* (2009), but produce inaccurate results on many real world images, as they only coarsely model the physics of image formation and ignore the global structure of the scene. None of the above approaches runs in real-time.

Global Priors Retinex-based methods have been extended to include non-local cues to improve the decomposition across an entire image (Gehler *et al.*, 2011; Shen & Yeo, 2011). Shen *et al.* (2008) and Zhao *et al.* (2012) show promising results for decomposing structured texture patterns by enforcing constant reflectance for pixels with similar local texture, but the non-local search is computationally expensive. Chang *et al.* (2014) present a probabilistic model for intrinsic decomposition. Other non-local methods enforce a small number of reflectance surfaces in the scene by clustering the reflectance image (Bi *et al.*, 2015; Garces *et al.*, 2012). Such complex clustering strategies are very time consuming and not real-time capable. This

approach includes non-local cues in a real-time capable way using a histogram-based clustering approach. Zoran *et al.* (2015) propose a framework to infer mid-level visual properties and apply it to the intrinsic decomposition task. Other computationally expensive global cues include creating pairwise pixel correspondences across the entire image (Bell *et al.*, 2014; Chen & Koltun, 2013). This method proposes similar correspondence constraints, which are real-time capable, through a non-local sampling strategy. In combination with the local sparsity term for reflectance, this method is able to achieve globally *and* temporally coherent decompositions.

Statistical and Learning-Based Techniques Statistics of real world geometry and illumination can be learned or modelled to help resolve the inherent ambiguity in intrinsic decomposition (Barron & Malik, 2015a). Such approaches are powerful, but often reach their limit on more complex scenes that fall outside of the used training data. Discriminative techniques have also been used to solve the Retinex problem by classifying edges as either a reflectance or shading edge (Bell & Freeman, 2001; Tappen *et al.*, 2005). Recently, Zhou *et al.* (2015) learned the relative reflectance ordering of image patches from a large annotated dataset to identify surfaces of similar reflectance under different illumination conditions. In spite of such diverse strategies, intrinsic decomposition remains a challenging, ill-posed problem, especially on real world scenes. Many recent approaches thus resort to user input like scribbles to resolve ambiguities (Bonneel *et al.*, 2014; Bousseau *et al.*, 2009; Shen *et al.*, 2011; Ye *et al.*, 2014). Even without such user-interaction, this approach produces decomposition results, in real-time, that are on par with or even better than results obtained with previous offline approaches.

Multi-Image and Depth-Based Techniques The highly under-constrained intrinsic decomposition problem benefits from additional information, such as per-pixel depth, temporal information from time lapses, or geometry from multi-view images. Several techniques rely on varying illumination over an image sequence of a static scene, to isolate the temporally constant reflectance from time-varying illumination effects (Hauagge *et al.*, 2013; Laffont & Bazin, 2015; Laffont *et al.*, 2012; Matsushita *et al.*, 2004; Weiss, 2001). Geometry cues computed from multi-view imagery are often exploited to construct further priors. Kong *et al.* (2014) use sequences captured with a moving light source, and use optical flow to find temporal correspondences in dynamic scenes. Surface normals are then used to improve the decompositions. Such approaches break down when lighting is near-constant, as in many real-life scenarios. Laffont *et al.* (2013) and Duchêne *et al.* (2015) use multi-view stereo to reconstruct

scene geometry and hence estimate environment maps of the scene. Depth information has proven very useful in estimating reflectance and shading, especially under a Lambertian reflectance assumption. Given an RGB-D video stream, illumination estimation and shape-from-shading refinement is feasible in real-time (Wu *et al.*, 2014). Depth information has also been exploited to impose local and global constraints on the shading layer (Barron & Malik, 2013; Chen & Koltun, 2013; Hachama *et al.*, 2015; Lee *et al.*, 2012), for example by exploiting local normal information. Although depth and other geometric cues are very valuable, they require specific multi-view capture, moving light sources or special camera hardware all of which are not available for live RGB video. The proposed method is the first approach for real-time, space-time coherent intrinsic decomposition from just a single monocular RGB video.

Intrinsic Video Decomposition Techniques Most discussed techniques are limited to decomposing a single image offline and yield unacceptable, temporally incoherent results when directly applied to video. Only few approaches explicitly tackle video. Shen *et al.* (2014) perform intrinsic decomposition only for specific regions in the video, their approach requires user input and has a slow offline runtime. Ye *et al.* (2014) propose a multi-pass optimization strategy for intrinsic video decomposition that clusters reflectance pixels and uses optical flow for correspondence across frames. Their approach is fundamentally offline as it takes more than a minute per video frame. Bonneel *et al.* (2015) use the temporal regularity of the input video as a guide to stabilize the shading and albedo layers computed by intrinsic decomposition techniques. Bonneel *et al.* (2014) suggest a fast and flexible method that uses both local and global chromaticity cues. However, since the method operates on grayscale images instead of RGB, the output reflectance image has the same chromaticity as the input image, which is often wrong. Therefore, the approach notably struggles if the assumptions of white light and Lambertian surfaces are violated. In contrast, the proposed method works in the RGB space and is more resilient against violation of these assumptions. The method of Bonneel *et al.* (2014) requires half a second per frame and an additional slow offline pre-processing step to calculate optical flow. In contrast, this approach runs completely in real-time. This method extends recent concepts for real-time non-linear optimization on the GPU (Wu *et al.*, 2014; Zollhfer *et al.*, 2014, 2015). In particular, a novel GPU-based optimizer is proposed to explicitly handle $\ell_2\ell_p$-optimization. Previous video techniques also use extensive user input, whereas this method obtains similar or even better results in real-time without any user-interaction.

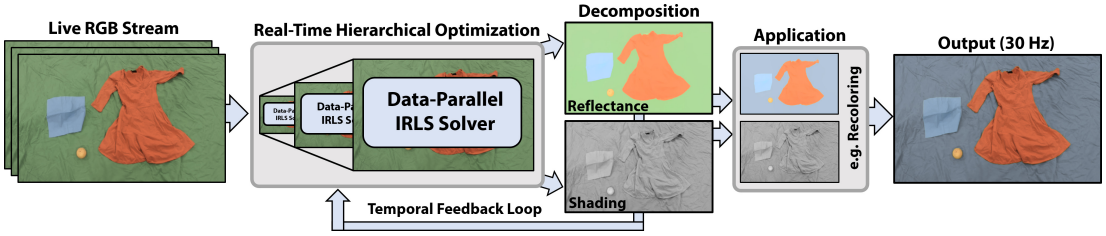


Figure 3.1: Overview of the proposed real-time intrinsic decomposition approach.

3.3 Overview

Given an arbitrary video stream as input, the proposed live intrinsic video decomposition technique extracts the corresponding shading and reflectance streams at real-time rates. Like previous decomposition methods, the proposed approach assumes Lambertian reflectance in the scene, i.e. the reflectance is equal to the albedo of the surface. Figure 3.1 shows an overview of all building blocks of the proposed approach. A novel mixed $\ell_2\ell_p$-formulation (see Section 3.4) is proposed that leads to decompositions that are both spatially and temporally coherent without the need for an explicit correspondence search. The resulting high-dimensional and non-convex variational optimization problem is robustly and efficiently optimized using a custom-tailored, fully data-parallel, iteratively reweighted least squares (IRLS) solver (see Section 3.5). Leveraging the computational power of modern graphics hardware, decomposition can be computed at frame rate. The obtained results (see Section 3.6) show that the proposed approach outperforms the state-of-the-art approaches qualitatively and quantitatively in terms of accuracy, robustness and runtime performance. The real-time capabilities of the proposed approach are shown in a live setup that demonstrates a variety of compelling demo applications (see Section 3.7), ranging from recolouring to material editing tasks. Finally, the theoretical and technical limitations (Section 3.8) are discussed and the chapter is concluded with an outlook (Section 3.9).

3.4 Energy

Intrinsic decomposition problems are commonly tackled by transferring and solving them in the log-domain (e.g. Shen & Yeo, 2011):

$$\mathbf{i}(\mathbf{x}) = \mathbf{r}(\mathbf{x}) + s(\mathbf{x}), \tag{3.2}$$

I is the input image, R is the reflectance image and S is the shading image. Lowercase letters are the log-domain versions of their upper-case counterparts. x is the

pixel location. This explicitly linearizes the constraints and facilitates the use of simpler optimization strategies. Even in the log-domain, the intrinsic decomposition problem is still under-constrained, since all per-pixel decompositions are completely independent. Most existing intrinsic video decomposition techniques rely on user scribbles to provide crucial constraints for solving the heavily under-constrained intrinsic decomposition problem. However, user scribbles are not an option for on-line intrinsic video decomposition approaches, as such user input cannot be provided at 30 Hz in a live-streaming setup. Previously used reflectance, shading and chromaticity priors are extended to suit the real-time setting. In addition, new global space-time and reflectance clustering priors are designed with real-time computational performance in mind, to solve the under-constrained decomposition problem. The approach is based on the decomposition energy

$$E(\mathcal{D}) = \sum_{\mathbf{x}} \Big[E_{\text{data}}(\mathbf{x}) + E_{\text{priors}}(\mathbf{x}) + E_{\text{non-local}}(\mathbf{x}) + E_{\text{clustering}}(\mathbf{x}) \Big]. \tag{3.3}$$

All sub-energies are defined per pixel $\mathbf{x}$. This energy is minimized for every video frame to obtain the decomposition

$$\mathcal{D} = \big[..., \mathbf{r}(\mathbf{x})^\top, ..., s(\mathbf{x}), ... \big]^\top \tag{3.4}$$

that stacks the unknown per-pixel reflectance and shading values defined by the vector-valued (RGB) reflectance layer $\mathbf{r}$ and the scalar shading layer s. All unknowns are defined in the log-domain. The image formation model in Equation (3.2) is assumed for defining the decomposition problem. Next, the particular data terms and prior constraints used in the novel decomposition energy are discussed and it is shown how to efficiently solve the resulting mixed $\ell_2\ell_p$-optimization problem at real-time rates. To this end, a specifically tailored data-parallel solution strategy is proposed in Section 3.5.

3.4.1 Data Fitting Term

The output of the optimization is a decomposition of the input video frame (in log-space) into a sum of reflectance and shading components. This is enforced as a soft-constraint via the data fitting term E_{data}. Similar to most previous intrinsic decomposition methods, monochromatic, white illumination is assumed; therefore the shading image is scalar-valued. In the log-domain,the fitting constraint is enforced per colour channel, i.e. $i_c \approx r_c + s$ for $c \in \{\text{R,G,B}\}$. To make the solution more robust to deviations from perfectly white illumination, per-channel perceptual

weights ω_c are applied to obtain the final constraint:

$$E_{\text{data}}(\mathbf{x}) = w_{\text{data}} \cdot \omega_{\text{iw}}(\mathbf{x}) \cdot \sum_{c \in \{\text{R,G,B}\}} \omega_c \cdot |i_c(\mathbf{x}) - r_c(\mathbf{x}) - s(\mathbf{x})|^2, \tag{3.5}$$

where $\{\omega_{\text{R}},\omega_{\text{G}},\omega_{\text{B}}\} = \{0.299,0.587,0.114\}$ (ITU-R BT.601). In addition, the data term is scaled by the data term weight w_{data}, and the image intensity weight

$$\omega_{\text{iw}}(\mathbf{x}) = 1 - w_{\text{intensity}} \cdot (1 - |\mathbf{I}(\mathbf{x})|), \tag{3.6}$$

which expresses the empirically confirmed observation that pixels with a higher intensity $|\mathbf{I}(\mathbf{x})|$ provide more reliable decomposition constraints, while low-intensity pixels need to be more strongly regularized to better deal with noise in the input data. In particular for commodity webcams, which have a low signal-to-noise ratio, low intensity pixels need strong regularization. This is adjustable via $w_{\text{intensity}}$.

3.4.2 Local Prior Terms

It is assumed that illumination effects such as shading and shadows only affect the intensity of a pixel, but not its chromaticity, $\mathbf{c}(\mathbf{x}) = \mathbf{I}(\mathbf{x})/|\mathbf{I}(\mathbf{x})|$. Therefore, any large gradient in the chromaticity does not originate in the shading image, but in the reflectance image. This can be interpreted as an intensity-normalized version of Retinex (Land & McCann, 1971). Based on a chromaticity similarity weight $\omega_{\text{cs}}(\mathbf{x},\mathbf{y})$, the reflectance and shading priors are selectively scaled, as described next, to compute an optimal decomposition.

$$\omega_{\text{cs}}(\mathbf{x},\mathbf{y}) = \exp(-\alpha_{\text{cs}} \cdot \|\mathbf{c}(\mathbf{x}) - \mathbf{c}(\mathbf{y})\|_2). \tag{3.7}$$

Here, the empirically determined factor $\alpha_{\text{cs}} = 15$ is used as it yields the best decomposition results in the experiments. In contrast to Bonneel *et al.* (2014), a smooth discriminator function is used instead of a hard threshold on chromaticity difference.

Reflectance Sparsity The reflectance image $\mathbf{r}$ is assumed to consist of piecewise-constant regions. Such a sparse solution can be obtained by minimizing the p^{th} power of the ℓ_p-norm, with $p \in [0,2)$, of the local per-pixel reflectance gradients $\nabla\mathbf{r}(\mathbf{x})$. Smaller choices of p yield sparser decompositions. The value $p = 0.8$ is set in all the experiments. However, as $\mathbf{r}$ is a 3-vector, $\nabla\mathbf{r}$ is a 3×2 matrix, consisting of horizontal and vertical gradients for each colour channel. To ensure soft and edge-friendly piecewise constancy of the reflectance image, ℓ_p-matrix norm is not

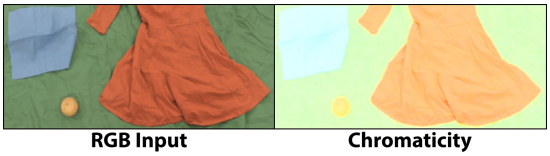


Figure 3.2: Chromaticity shift: in practical conditions, especially in dark regions (e.g. folds of the dress), chromaticity changes occur due to indirect illumination effects and finite camera sensitivity.

minimized directly, but instead the gradients are separated along each dimension and their magnitude minimized independently:

$$E_{\text{reflectance}}(\mathbf{x}) = w_{\text{reflectance}} \cdot \sum_{\mathbf{y} \in N(\mathbf{x})} \omega_{\text{cs}}(\mathbf{x},\mathbf{y}) \cdot \|\mathbf{r}(\mathbf{x}) - \mathbf{r}(\mathbf{y})\|_2^p. \tag{3.8}$$

Here, $N(\mathbf{x})$ is the 4-pixel neighbourhood of pixel $\mathbf{x}$, and the more similar two pixels' chromaticities, as measured by $\omega_{\text{cs}}(\mathbf{x},\mathbf{y})$, the lower the weight on the reflectance difference. The whole objective is scaled by $w_{\text{reflectance}}$. Note that this constraint is expressed directly on colour values, not on gradients (Bonneel *et al.*, 2014), which benefits real-time performance (see Section 3.6.5).

Shading Smoothness For purely diffuse surfaces, shading is only a function of the shape of the object. Since objects in natural scenes generally have smooth shapes, the shading image is also expected to be smooth. In addition, neighbouring pixels with different chromaticities, as measured by $1 - \omega_{\text{cs}}(\mathbf{x},\mathbf{y})$, indicate a reflectance edge, where shading smoothness should be more strongly enforced:

$$E_{\text{shading}}(\mathbf{x}) = w_{\text{shading}} \cdot \sum_{\mathbf{y} \in N(\mathbf{x})} (1 - \omega_{\text{cs}}(\mathbf{x},\mathbf{y})) \cdot |s(\mathbf{x}) - s(\mathbf{y})|^2. \tag{3.9}$$

Here, w_{shading} is the weight of this prior constraint.

Chromaticity Prior As mentioned earlier, it is assumed that the chromaticity of the input image is not altered by illumination effects such as shading and shadows. In this case, the chromaticity of the unknown reflectance image $\mathbf{r}$ should be the same as that of the input image. This is enforced using the soft constraint

$$E_{\text{chromaticity}}(\mathbf{x}) = w_{\text{chromaticity}} \cdot \|\mathbf{c}(\mathbf{x}) - \mathbf{c}_{\text{r}}(\mathbf{x})\|_2^2, \tag{3.10}$$

where $\mathbf{c}$ is the chromaticity of the input video frame, and $\mathbf{c}_{\text{r}}$ is the chromaticity of the reflectance image $\mathbf{r}$.

In a simplified image formation model that only considers direct white illumination and infinite camera precision, chromaticity changes solely occur due to reflectance changes. However, in the real world (especially in low-intensity regions),

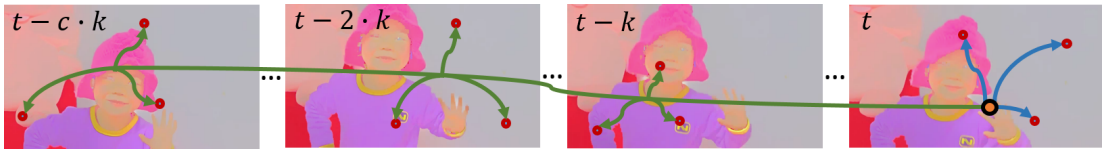


Figure 3.3: Spatio-temporal reflectance consistency prior: global consistency constraints are applied in the space (blue) and time (green) domains based on random sampling. If sampled pixels have similar chromaticity, their reflectances are constrained to also be similar.

indirect illumination effects and the camera's finite sensitivity limit this assumption. This leads to shifts in the captured chromaticities (see Figure 3.2). In brighter regions, the chromaticity is still a good approximation of the reflectance. Therefore, the three priors are combined using the image intensity weight $\omega_{\text{iw}}(\mathbf{x})$, to reduce the influence of the shading and chromaticity priors for dark pixels, to obtain

$$E_{\text{priors}}(\mathbf{x}) = E_{\text{reflectance}}(\mathbf{x}) + \omega_{\text{iw}}(\mathbf{x}) \cdot \left[E_{\text{shading}}(\mathbf{x}) + E_{\text{chromaticity}}(\mathbf{x})\right]. \tag{3.11}$$

3.4.3 Spatio-Temporal Reflectance Consistency Prior

Many natural and man-made scenes contain multiple, identically coloured instances of an object, such as cushions on a sofa. Illumination is also changing over time, causing pixels to increase or decrease in brightness. In these scenarios, it is essential to ensure spatio-temporally consistent reflectances. This is not handled by the constraints described so far, which merely locally enforce piecewise constant reflectance. To ensure spatially and temporally consistent reflectance, a new global, sampling-based, spatio-temporal reflectance consistency constraint is proposed, that does not rely on costly space-time correspondence finding, such as optical flow. This allows for real-time performance.

Each pixel $\mathbf{x}$ in the reflectance image is connected to N_{s} randomly sampled pixels $\mathbf{y}_i$. Samples are chosen from reflectance images of the current and previous frames t_i, as illustrated in Figure 3.3. If the chromaticity of the current pixel is reasonably close to that of the sampled pixel, their reflectances are constrained to be similar:

$$E_{\text{non-local}}(\mathbf{x}) = w_{\text{non-local}} \cdot \sum_{i=1}^{N_{\text{s}}} g_i(\mathbf{x}) \cdot \left\|\mathbf{r}(\mathbf{x}) - \mathbf{r}_{t_i}(\mathbf{y}_i)\right\|_2^2 \tag{3.12}$$

$$g_i(\mathbf{x}) = \left\{ \begin{array}{ll} \omega_{\text{iw}}(\mathbf{x}) & \text{if } \left\|\mathbf{c}(\mathbf{x}) - \mathbf{c}_{t_i}(\mathbf{y}_i)\right\|_2 < \tau_{\text{cc}} \\ 0 & \text{otherwise.} \end{array} \right\} \tag{3.13}$$

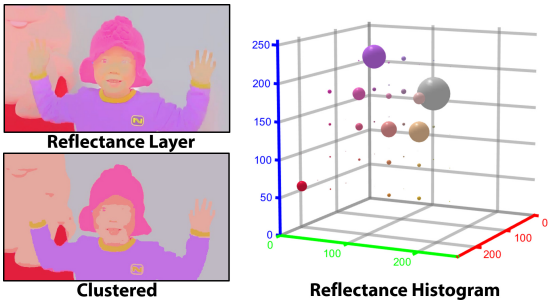


Figure 3.4: Reflectance Clustering: The reflectance layer is clustered based on a weighted k-means strategy on the reflectance histogram.

Here, τ_{cc} is a chromaticity consistency threshold. $N_{\text{s}} = 9$ pixel locations are randomly sampled from the current frame t as well as the previous five keyframes (spaced five frames apart). Since darker pixels suffer from shifted chromaticities, their contribution is again reduced based on ω_{iw}.

The proposed approach, although relying on random sampling, is especially effective when combined with the reflectance sparsity prior. It is very likely that distinct regions of same reflectance are connected by at least a few samples, and the reflectance sparsity prior then spreads the global reflectance consistency constraints to other nearby pixel locations. By creating connections to previous video frames, this term leads to temporally stable decompositions. The number and spacing of the used frames is adjustable: a shorter temporal window may for example be preferable in case of fast motion or illumination changes. Spacing the frames further apart makes the approach more resilient to slow illumination changes. A default of five past keyframes spaced five frames apart was used, which proved to be sufficient for all the test sequences.

3.4.4 Reflectance Clustering Prior

The reflectance sparsity and non-local consistency priors lead very close to the goal of a sparse distribution of reflectances, by encouraging piecewise constancy and consistent colours for disjoint objects of the same reflectance, respectively. However, there may still be remaining inconsistencies in actually uniform reflectance regions and unwanted temporal changes within the same material. Therefore, a per-pixel soft constraint for global reflectance consistency is introduced that ensures the reflectance image to be close to the desired result and temporally stable, even without costly spatial correspondence finding. This is achieved by estimating a clustered version of the reflectance image. First, a histogram of the reflectance image is computed and major reflectance clusters found. Each pixel's reflectance is then constrained to match the reflectance of its most similar cluster. Specifically,

an RGB histogram of the reflectance image is computed with 30^3 uniformly spaced bins, where each bin stores the number of pixels within it, as well as their mean colour (see Figure 3.4). Histograms are exponentially averaged over time to improve the temporal coherence of the reflectance clusters, by performing weighted k-means clustering on the reflectance histogram. The cluster centers are initialized with the previous frame's clusters, which speeds up convergence, or randomly in the case of the first frame. Duplicate reflectance clusters with chromaticity differences below the chromaticity consistency threshold τ_{cc} are collapsed before use.

A clustered reflectance image $\mathbf{r}_{\text{cluster}}$ is then created using the closest reflectance cluster for each reflectance pixel $\mathbf{r}(\mathbf{x})$ in terms of ℓ_2 distance. This clustered reflectance image could be used directly as the final reflectance image, but any errors in the clustering process would become part of the final result. Instead,the clustered reflectance image is used as a soft constraint that is most strongly applied to dark pixels as these are most unreliable. The reason for this is what is named here as *chromaticity shift*: large shading variations may cause a shift in chromaticity in the darker regions of the same reflectance surface because of inter-reflections and finite camera sensitivity. This issue is resolved by constraining dark pixels more strongly to be similar to their closest reflectance cluster:

$$E_{\text{clustering}}(\mathbf{x}) = \omega_{\text{clustering}}(\mathbf{x}) \cdot \|\mathbf{r}(\mathbf{x}) - \mathbf{r}_{\text{cluster}}(\mathbf{x})\|_2^2, \tag{3.14}$$

$$\omega_{\text{clustering}}(\mathbf{x}) = w_{\text{clustering}} \cdot \exp(-\alpha_{\text{clustering}} \cdot |\mathbf{I}(\mathbf{x})|), \tag{3.15}$$

using the clustering prior weight $w_{\text{clustering}}$ and empirically determined soft function constant $\alpha_{\text{clustering}} = 0.4$. Using the clustered reflectance image to define the decomposition energy is a chicken-and-egg problem, as estimating the clustered image requires the reflectance to be available, whereas estimating the reflectance requires the clustering. To solve this problem, the coarse-to-fine optimization strategy is exploited (see Section 5.4). The clustering is performed on the reflectance estimated on the second-finest level and used for regularizing the finest level result.

3.5 Optimization

The intrinsic decomposition objective $E(\mathcal{D}) : \mathbb{R}^{4N} \to \mathbb{R}$ proposed in Equation (3.3) is a mixed $\ell_2\ell_p$-optimization problem in the unknown parameter values $\mathcal{D}$. Here, $N = W \times H$ is the resolution of the input video stream. The parameter vector $\mathcal{D}$ holds the $4N$ unknown pixel values that fully define the intrinsic decomposition,

i.e. the per-pixel log-space reflectance $\mathbf{r}(\mathbf{x}) \in \mathbb{R}^3$ and shading $s(\mathbf{x}) \in \mathbb{R}$. The optimal decomposition $\mathcal{D}^*$ is the minimizer of $E(\mathcal{D})$:

$$\mathcal{D}^* = \underset{\mathcal{D}}{\operatorname{argmin}} E(\mathcal{D}). \tag{3.16}$$

This high-dimensional, under-constrained optimization problem is non-linear and non-convex due to the involved ℓ_p-optimization. In addition, this optimization has a large number of unknowns even for small video resolutions, e.g. about 2 million unknowns for a resolution of 800×600 pixels, which have to be optimized under the tight real-time constraint of 30 Hz. Previously, sparse gradient priors (Bonneel *et al.*, 2014; Joshi *et al.*, 2009; Levin & Weiss, 2007; Levin *et al.*, 2007) have been tackled on the CPU using an *iteratively reweighted least squares* (IRLS) approach; but not at real-time rates given millions of unknowns. Here, the computational horsepower of the data-parallel GPU architecture is used to solve such variational optimization problems at frame rate. In contrast to previous work on data-parallel optimization (Wu *et al.*, 2014; Zollhfer *et al.*, 2014, 2015), which only deals with standard non-linear least squares formulations, a novel solution strategy is proposed for general unconstrained ℓ_p-optimization problems. To this end, a custom-tailored data-parallel IRLS solver is devised that allows to solve for up to 2 million unknowns at real-time rates.

3.5.1 Data-Parallel IRLS Core Solver

IRLS is a widely used optimization strategy (Holland & Welsch, 1977); its key idea is to transform a general unconstrained optimization problem to a sequence of reweighted subproblems:

$$\left\{ \mathcal{D}^{(k)} = \underset{\mathcal{D}}{\operatorname{argmin}} E^{(k)}\left(\mathcal{D} \,|\, \mathcal{D}^{(k-1)}\right) \right\}_{k=1}^{K}. \tag{3.17}$$

The original energy E is successively reweighted based on the previous solution $\mathcal{D}^{(k-1)}$ to obtain new energies $E^{(k)}$. Starting from an initial estimate $\mathcal{D}^{(0)}$, the optimum $\mathcal{D}^* = \mathcal{D}^{(K)}$ of E is found based on K such steps. For the first time, the IRLS strategy is integrated into a data-parallel iterative GPU solver for handling the ℓ_p term in the energy. As a starting point, consider a single scalar ℓ_p-residual of the objective. Since the p^{th} power of ℓ_p is used in the energy, it can be written as:

$$\left| r(\mathcal{D}^{(k)}) \right|^p. \tag{3.18}$$

Here, $r(\mathcal{D}^{(k)}) \in \mathbb{R}$ is a general scalar and linear residual. Now let $\mathcal{D}^{(k-1)}$ be the approximate solution computed in the previous iteration step. Then, a suitable reweighting scheme is obtained by approximately splitting Equation (3.18) into two components:

$$\left|r(\mathcal{D}^{(k)})\right|^p \approx \underbrace{\left|r(\mathcal{D}^{(k-1)})\right|^{p-2}}_{c(\mathcal{D}^{(k-1)})} \cdot \left|r(\mathcal{D}^{(k)})\right|^2 \tag{3.19}$$

$$= \left(\sqrt{c\big(\mathcal{D}^{(k-1)}\big)} \cdot r(\mathcal{D}^{(k)})\right)^2 . \tag{3.20}$$

This factorization is based on the assumption that parameters change slowly $\mathcal{D}^{(k)} \approx \mathcal{D}^{(k-1)}$. The reweighting factor $c\big(\mathcal{D}^{(k-1)}\big)$ is constant during one iteration step, since it only depends on the previous solution. The remaining second factor is a quadratic function of the parameters since the residuals $r(\mathcal{D}^{(k)})$ are linear. Note, reweighting also applies to the case $p=2$, resulting in $c\big(\mathcal{D}^{(k-1)}\big)=1$. Thus,the energy $E^{(k)}$ can be written using reweighting factors $c_k(\mathcal{D}^{(k-1)})$:

$$E^{(k)}(\mathcal{D} \,|\, \mathcal{D}^{(k-1)}) = \sum_{k=1}^{M} \Bigg(\underbrace{\sqrt{c_k(\mathcal{D}^{(k-1)})} \cdot r_k(\mathcal{D})}_{\hat{r}_k(\mathcal{D}|\mathcal{D}^{(k-1)})} \Bigg)^2 . \tag{3.21}$$

The total number $M = N(13+N_s)$ of residuals $\hat{r}_k(\mathcal{D} \,|\, \mathcal{D}^{(k-1)})$ depends on the data fitting term ($3N$ terms), shading smoothness prior (N terms), reflectance sparsity prior ($3N$ terms), chromaticity prior ($3N$ terms), spatio-temporal reflectance coherence prior (NN_s terms) and the reflectance clustering prior ($3N$ terms). To simplify notation further, all M scalar residual terms $\hat{r}_k(\mathcal{D} \,|\, \mathcal{D}^{(k-1)})$ are stacked in a single vector:

$$F^{(k)}(\mathcal{D} \,|\, \mathcal{D}^{(k-1)}) = \left[\hat{r}_1(\mathcal{D} \,|\, \mathcal{D}^{(k-1)}), \ldots, \hat{r}_M(\mathcal{D} \,|\, \mathcal{D}^{(k-1)})\right]^\top . \tag{3.22}$$

This vector can be interpreted as a high-dimensional vector field $F : \mathbb{R}^N \to \mathbb{R}^M$ that allows to rewrite $E^{(k)}(\mathcal{D})$:

$$E^{(k)}(\mathcal{D} \,|\, \mathcal{D}^{(k-1)}) = \left\| F^{(k)}(\mathcal{D} \,|\, \mathcal{D}^{(k-1)}) \right\|^2 . \tag{3.23}$$

Since all elements of $F^{(k)}$ are linear functions of the unknowns, the resulting optimization problem is quadratic, hence convex:

$$\mathcal{D}^{(k)} = \underset{\mathcal{D}}{\operatorname{argmin}} \left\| F^{(k)}\big(\mathcal{D} \,|\, \mathcal{D}^{(k-1)}\big) \right\|_2^2 . \tag{3.24}$$

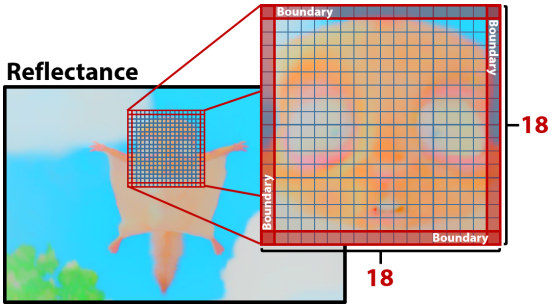


Figure 3.5: Subdomains of the localglobal optimization approach.

The global optimum of the sequential sub-problems is found by setting the partial derivatives to zero. The resulting highly over-constrained linear system ($M \gg N$) is solved in the least-squares sense. Previous work (Weber *et al.*, 2013; Wu *et al.*, 2014; Zollhfer *et al.*, 2014, 2015) demonstrated the feasibility of data-parallel preconditioned conjugate gradient (PCG) for the fast solution of such problems. A similar GPU-based PCG approach is used to exploit the sparsity pattern of the system matrix. Entries of the system matrix are computed on the fly (and only if they are required) during PCG iterations, and are never explicitly stored. As a preconditioner, inverse diagonal preconditioning is employed. The proposed strategy is highly efficient and already provides real-time performance for a moderate amount of unknowns. However, since the objective has millions of unknown parameters, real-time optimization is not directly feasible with the proposed core solver. To alleviate this problem, a local-global optimization approach is proposed that exploits the regular grid structure of the image domain to partition the problem into small local sub-problems. Each small sub-problem can then be solved efficiently in shared GPU memory based on the presented core solver.

3.5.2 LocalGlobal Optimization Approach

Instead of solving the global joint optimization problem directly, the domain is subdivided into small square subdomains and the optimization is performed locally on each of these. Afterwards, the updates obtained in this local step are exchanged, and the whole procedure is iterated. For a start, consider the energy without the global reflectance consistency constraint. A strategy to incorporate this energy term is described later in Section 3.5.3. The evaluation of all other objectives requires locally at most a one-ring pixel neighbourhood. Each sub-problem is solved independently by one thread block on the GPU and it is aimed to keep the complete state of the solver close to the associated multiprocessor, i.e. in shared memory and registers.

In each subdomain, input data and current decomposition is first cached to shared memory. In this step, a one-ring boundary is included. Neumann constraints are enforced on this boundary to decouple the sub-problems. The size of the local subdomains is set based on the available L1 cache on the used GPU. 16×16 subdomains are used, see Figure 3.5. Including the boundary pixels, this leads to overlapping 18×18 regions that are loaded to shared memory. The local per-domain problem is solved via the proposed IRLS strategy. After solving the local problems, the subdomain decomposition result is written back to global memory to facilitate data exchange between regions. For the 16×16 inner subregions, one thread per pixel writes the obtained new shading and reflectance values to global memory. Values on the boundary are not written back, as they are part of the inner subregion of an adjacent subdomain. This can be interpreted as a variant of the *Schwarz Alternating Procedure* (Zhao, 1996) for domain decomposition problems. Note that in the implementation, IRLS steps and Schwarz iterations are directly interleaved. The global memory is written to out-of-place, leading to deterministic results (fully *Additive Schwarz*), which are independent of GPU scheduling. This is in contrast to Wu *et al.* (2014) and Zollhfer *et al.* (2015), where a blend between an *Additive* and *Multiplicative* strategy has been proposed. This approach leads to temporally more coherent results if only a fixed limited number of iterations is performed. Sub-domains are shifted virtually after each iteration step based on a Halton sequence to improve convergence.

3.5.3 Adding the Spatio-Temporal Reflectance Prior

Up to now, the spatio-temporal reflectance prior was not considered in the optimization strategy. This energy term does not directly fit the proposed localglobal sub-domain optimization strategy due to its global nature, since sample points are randomly distributed in the video volume. This introduces a coupling between the local subproblems. Note that the optimization strategy proposed by Wu *et al.* (2014) and Zollhfer *et al.* (2015) can not handle this situation. A two-fold strategy is followed to deal with this problem. First, these connections are treated similar to the boundary by imposing Neumann constraints for values outside of the processed sub-domain. This allows to cache these values dynamically to registers before the local sub-domain optimization commences. Second, unidirectionality of the constraints is assumed, i.e. only the reflectance value at the currently processed pixel $\mathbf{r}(\mathbf{x})$ in Equation (3.12) is an unknown and the target $\mathbf{r}_{t_i}(\mathbf{y}_i)$ is assumed to be constant. Informally speaking, pixels only see their drawn samples, but do not know if they have been sampled by others. Therefore, the partial derivatives do

not depend on the target, and a constant amount of N_s values per thread has to be cached. these values are kept in registers. Cached values are updated over the non-linear IRLS iterations. This decouples the sub-domain systems from each other and allows for a data-parallel optimization as proposed earlier.

3.5.4 Nested Hierarchical Optimization

For the solution strategy proposed so far, error reduction stalls after the high-frequency error components have been resolved. Low frequency errors are only slowly resolved, since the propagation of updates over long spatial distances requires many iteration steps. This is a common problem of all iterative solution strategies. To alleviate this problem, the proposed iterative localglobal optimization approach is run in a nested coarse-to-fine loop based on a Gaussian pyramid. Since low frequency errors are of higher frequency on the coarser resolution levels, all frequency components of the error can be efficiently handled, hence leading to fast convergence. The optimization is solved on every level and a prolongation operator is used to obtain a suitable starting value for the next finer level. Prolongation is based on bi-linear interpolation of pixel data. Currently, a hierarchy with three to four levels is used depending on the input resolution. This turns out to be sufficient for good convergence rates. On the coarsest level, a frame-by-frame initialization is performed based on the assumption that reflectance and shading have the same magnitude. Therefore, $\mathbf{r}(\mathbf{x}) = \mathbf{i}(\mathbf{x})/2$ and $s(x) = |\mathbf{i}(\mathbf{x})|/2$ are set. The reflectance clustering prior (Section 3.4.4) is applied only on the finest pyramid level, and reflectance image computed on the second-finest pyramid level is used to compute the reflectance clusters.

3.6 Experiments

The approach was tested on several challenging real and synthetic datasets to evaluate its robustness, accuracy and runtime behaviour in comparison to the state-of-the-art. The test datasets consist of some real sequences (GIRL[1], TOY[1], DOWNSTAIRS[1], OBJECTS[1], HOUSE[2], CART[2]) and some synthetic sequences (SQUIRREL[1], CATHEDRAL[1], SANMIGUEL[2]) provided by existing intrinsic video decomposition techniques. In addition, the approach is applied to several new live video streams captured by a webcam for demonstrating various applications. Both qualitative and quantitative analysis of the results are performed in comparison to the intrinsic video decomposition

[1] http://media.au.tsinghua.edu.cn/yegenzhi/IntrinsicVideo.htm
[2] http://liris.cnrs.fr/~nbonneel/intrinsic.htm

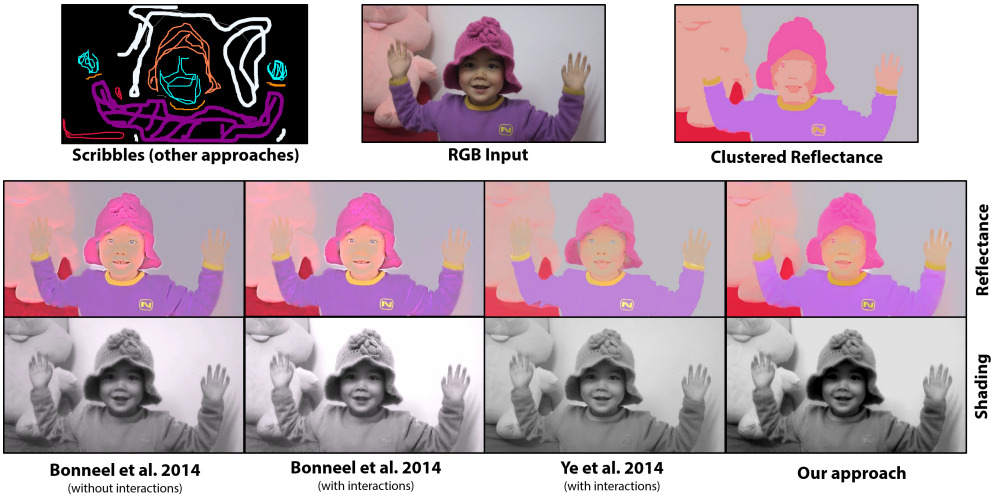


Figure 3.6: State-of-the-art comparison to Bonneel *et al.* (2014) and Ye *et al.* (2014) on the GIRL sequence. This approach obtains comparable or even higher-quality decompositions than previous approaches (less shading in the reflectance layer), while being orders of magnitude faster (10× faster than Bonneel *et al.* and 1800× faster than Ye *et al.*) and not requiring user input in the form of scribbles.

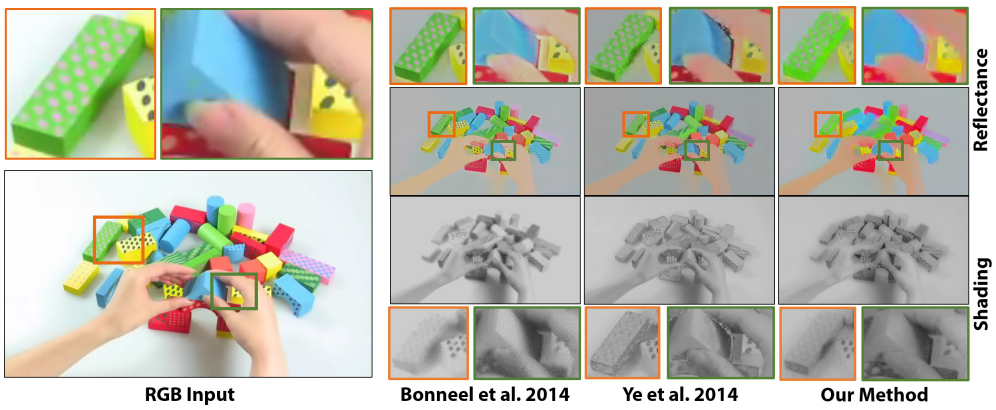


Figure 3.7: State-of-the-art comparison to Bonneel *et al.* (2014) and Ye *et al.* (2014) on the TOY sequence. The approach obtains decompositions of higher quality than previous approaches (less shading in the reflectance layer, sharper shading layer, less artifacts), while being orders of magnitude faster (10× faster than Bonneel *et al.*, 200× faster than Ye *et al.*) and not requiring user input in the form of scribbles.

methods of Bonneel *et al.* (2014) and Ye *et al.* (2014). This method deals better with illumination effects such as shadows and shading than previous approaches, while being orders of magnitude faster. In the quantitative comparisons, smaller decomposition errors than current state-of-the-art video techniques are consistently obtained.

In most experiments, the following fixed set of parameters are used to instantiate the intrinsic decomposition energy: $w_{\text{reflectance}} = 0.5$, $w_{\text{intensity}} = p = 0.8$, $w_{\text{cs}} = 1$, and

$w_{\text{data}} = w_{\text{shading}} = w_{\text{chromaticity}} = w_{\text{non-local}} = w_{\text{clustering}} = 10$. Note that this approach works out of the box for all sequences evaluated here, with resolutions ranging from 640×360 to 960×540, including the live video footage in the same resolution range. Drastic deviation from this range may require parameter adjustments. Since the intrinsic decomposition problem is ambiguous, the reflectance layer is globally scaled based on a single scalar (the shading layer is scaled inversely) to match the perceived brightness of previous state-of-the-art approaches. Note, the scaled results are still valid decompositions.

3.6.1 Qualitative Evaluation

The first qualitative comparison is with the state-of-the-art approaches of Bonneel *et al.* (2014) and Ye *et al.* (2014) in Figure 3.6. This approach obtains reflectance layers of higher quality, particularly in the more uniform regions (see the hat). The other two approaches more strongly bake shading variation into the reflectance map. The input (see creases of the shirt) is decomposed better into its reflectance and shading components. This is possible due to the novel spatio-temporal prior. Note, the other methods require intricate user-interaction, in the form of constant reflectance scribbles in the first frame of the video, to obtain reasonable decomposition results, whereas this approach is fully automatic and orders of magnitude faster (10× faster than Bonneel *et al.*, 1800× faster than Ye *et al.*). In addition, the method of Bonneel *et al.* (2014) operates on grayscale images instead of RGB data. Therefore, the output reflectance has the same chromaticity as the input. This leads to artifacts if the assumption of white light or Lambertian surface is violated.

The global spatio-temporal prior ensures that reflectance values of spatially or temporally distant objects with the same appearance are similar in the decomposition. This becomes especially apparent in the TOY sequence (see Figure 3.7), which contains several toy blocks with similar appearance. The previous state-of-the-art approaches struggle with this challenging scenario. In particular, they are unable to uniformly decompose the blue coloured blocks and lead to a lot of shading detail becoming part of the reflectance layer. Note again, this method is orders of magnitude faster than these approaches and does not require user input in the form of scribbles.

3.6.2 Quantitative Evaluation

Established error metrics (Grosse *et al.*, 2009) are used to compare the results to ground-truth data:

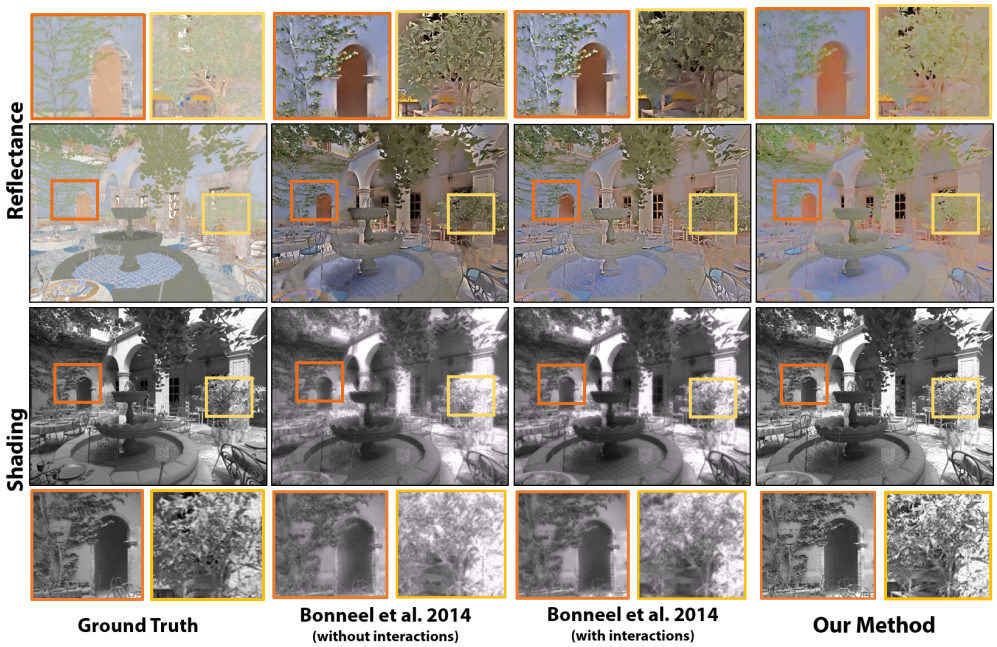


Figure 3.8: ground-truth comparison on the SANMIGUEL sequence. This approach obtains decompositions that more closely match the ground-truth. Bonneel *et al.*'s result artificially blurs the shading layer and contains small-scale shading in the reflectance layer. Even user-provided scribbles do not alleviate this issue. This approach is also one order of magnitude faster and can be applied to live video data.

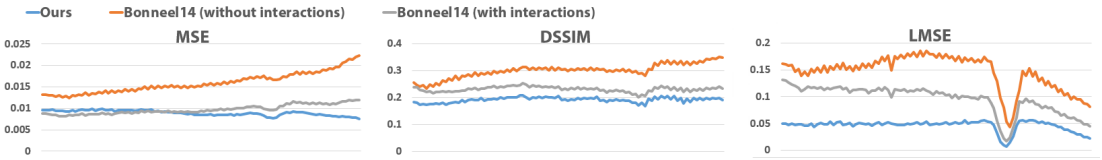


Figure 3.9: Quantitative evaluation: this approach obtains lower MSE, DSSIM and LMSE errors than the approach of Bonneel *et al.* (2014) on the SANMIGUEL sequence, while also being one order of magnitude faster and not relying on user input in the form of scribbles.

1. MSE (*mean squared error*) measures the average of the squared per-pixel deviations from the ground-truth. In case of colour images, it is averaged over all channels.

2. LMSE (*local mean squared error*) measures the average MSE over a set of overlapping patches. The intensity of each patch is scaled by a single scalar value to minimize the per-patch MSE value. The metric is normalized so that an estimate of all zeros has the maximum possible score of 1. A patch size of 10×10 is used.

	MSE			LMSE			DSSIM		
Approach	shading	reflectance	mean	shading	reflectance	mean	shading	reflectance	mean
Bonneel *et al.* (2014) (no scribbles)	0.0063	0.0258	0.0161	0.1564	0.1332	0.1447	0.2794	0.3226	0.3011
Bonneel *et al.* (2014) (scribbles)	0.0030	0.0166	0.0097	0.0886	0.1029	0.0947	0.1753	0.2898	0.2302
The proposed approach	0.0028	**0.0151**	**0.0089**	**0.0309**	**0.0622**	**0.0461**	**0.1304**	**0.2566**	**0.1915**
The proposed approach (w/o non-local prior)	**0.0027**	0.0173	0.0099	0.0421	0.0961	0.0688	0.1367	0.2693	0.2014

Table 3.1: Quantitative comparison on the SANMIGUEL sequence: The obtained decompositions have a lower error (bold) than previous work.

3. DSSIM (*structural dissimilarity index*) is an information theoretic metric that measures the perceived change in structural information between two images.

Each metric is computed separately for the reflectance and shading images, and also the average is reported as a final result.

Figure 3.8 compares the results on the synthetic SANMIGUEL sequence to the approach of Bonneel *et al.* (2014). The proposed approach achieves higher quality decompositions, especially in the foliage and on the background walls. The complex illumination pattern on the leaves is difficult to decompose with previous state-of-the-art approaches, even with user-interaction in the form of scribbles. This approach is able to obtain decompositions of better quality fully automatically even in this challenging scenario. Note, this approach is also an order of magnitude faster. Figure 3.9 shows the per-frame MSE, LMSE and DSSIM results as plots for the complete sequence. This approach obtains consistently lower decomposition errors in almost all frames of the sequence. The increased temporal stability of this approach, compared to Bonneel *et al.* (2014), can be seen in the smaller variance of the error plots. The errors over the complete sequence are summarized in Table 3.1, separately for shading and reflectance layers, and averaged, and also indicate the superior performance of this approach, even without using user scribbles.

3.6.3 Evaluation on 'Intrinsic Images in the Wild' Dataset

The approach is additionally evaluated on the 'Intrinsic Images in the Wild' benchmark dataset of Bell *et al.* (2014) containing static images. Towards this goal, the temporal consistency prior term is disabled, the 5,230 individual images in the dataset are decomposed and the *weighted human disagreement rate* (WHDR) evaluated, which compares the manual annotations on the images with the decomposed reflectance images. A $\text{WHDR}_{10\%}$ score of 31.4% is obtained. Note that this technique is not meant to compete with traditional intrinsic *single-image* decomposition techniques, as a different set of challenges in intrinsic decomposition of *live videos* are addressed here.

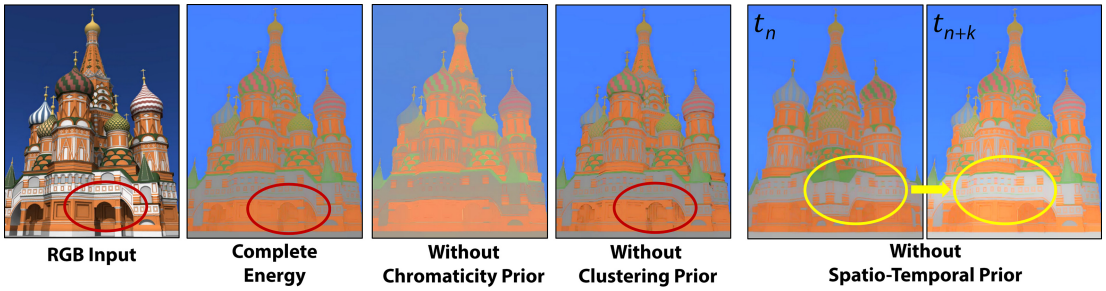


Figure 3.10: Influence of energy terms: reflectance result on the CATHEDRAL sequence. The best reflectance image is obtained with the full energy. Without the **chromaticity prior**, the output reflectance colour deviates from the input. The **clustering prior** removes shading variation from the reflectance layer (red circles). Without the **spatio-temporal prior**, the decomposition is temporally unstable (yellow circles).

Figure 3.11: Influence of different priors on the SANMIGUEL sequence. The sampling-based global spatial prior constraint helps to remove shading variations from the reflectance layer.

3.6.4 Influence of the Different Energy Terms

This intrinsic decomposition approach obtains high-quality results due to the carefully crafted decomposition energy function. Next, the relative importance of the different objective terms is evaluated. Figure 3.10 shows the reflectance images for different instantiations of the decomposition energy, where certain components are disabled by setting the respective weight to zero. The best decomposition results are obtained by the full combined energy. The chromaticity prior helps to keep the output reflectance close to the input's chromaticity leading to more saturated results. The clustering prior is particularly useful in decomposing the challenging dark shadow regions in the image accurately. Without it, illumination effects such as shadows and shading become part of the reflectance layer. The spatio-temporal prior ensures the global consistency of the reflectance layer, even for disconnected regions of the same material. In addition, it leads to temporally coherent results.

Sequence	Resolution	Time
HOUSE	1024 × 576	36.0 ms
GIRL	960 × 540	31.8 ms
DOWNSTAIRS	960 × 540	31.5 ms
TOY	640 × 360	16.1 ms
SQUIRREL	854 × 480	26.0 ms
SANMIGUEL	1280 × 960	68.6 ms
Live	640 × 480	22.1 ms

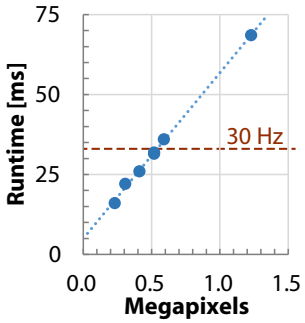


Table 3.2: Runtime performance for different input resolutions.

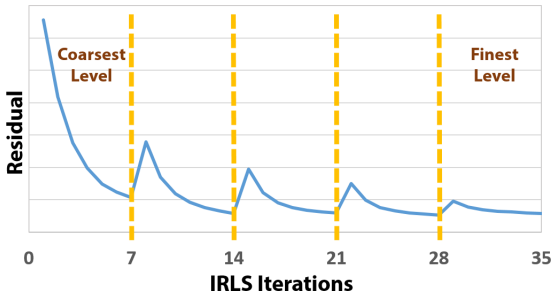


Figure 3.12: Convergence: The residual error is always decreasing.

The added global spatial consistency can even better be judged from the SANMIGUEL sequence (see Figure 3.11). Note that the background wall in the courtyard, the floor and the leaves, all incorrectly contain illumination and shadows if this prior is not applied. The lower error in the ground-truth comparison (see Table 3.1) also reflects this difference in quality. Therefore, all proposed priors contribute significantly to the accuracy of the obtained decomposition results.

3.6.5 Runtime and Convergence

Figure 3.12 shows the convergence behaviour of the novel nested IRLS approach. The staircase pattern corresponds to the number of hierarchy levels (5 in this case). For this experiment, 7 non-linear IRLS iterations were used per level with 8 PCG steps each. As can be seen, the IRLS approach converges on each hierarchy level in about 4 iteration steps. Due to the used hierarchy, global convergence is fast and all error frequencies are efficiently resolved. Since convergence on a single level is reached after only a few iteration steps, in the following, the number of IRLS iterations is set to 4; all other settings are kept unchanged. This is a good trade-off between accuracy and runtime performance. The mean per-frame runtime of the approach is given for

Figure 3.13: Reflectance recolouring on the GIRL sequence. The girl's is recoloured in real-time using this intrinsic decomposition approach. Note that the shading detail is preserved.

seven sequences with different input resolutions in Table 3.2. Runtime is essentially linear in the number of pixels in the video, and frame rates of more than 30 Hz are achieved for input resolutions up to 950×540. In particular, live sequences at VGA resolution are processed in less than 23 ms, which guarantees real-time feedback. All timings have been measured on a commodity Nvidia GTX Titan graphics card.

3.7 Applications

This approach, for the first time, enables high-quality intrinsic decompositions in real-time. This real-time capability is the basis for a large variety of video editing applications, which are showcased in a live setup. The live setup is based on a commodity webcam (*Logitech HD Pro C920*), which captures RGB video at 30 Hz. A colour resolution of 640×480 is used for all applications. The camera's exposure, white balance and focal length were manually set to a fixed value.

3.7.1 Dynamic Reflectance Recolouring

This demo showcases the realistic recolouring of different objects in live video footage. For each captured frame, the intrinsic decomposition is first computed and then chromaticity keying is applied to the reflectance layer to select a subregion for which a different reflectance value is set. Note that in the recoloured composite (see Figure 3.13), shading variations are realistically preserved. The real-time setting enables immediate visual feedback, even if parameters are changed.

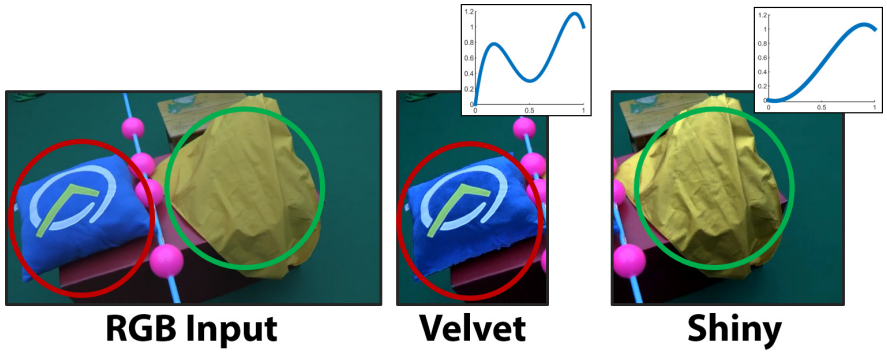


Figure 3.14: Editing material appearances on the OBJECTS sequence. The cushion looks like velvet (red circles), and the cloth is modified to appear shinier (green circles). The blue curves show the tone mapping applied to the corresponding regions of the shading layer to achieve the effect in each case.

Figure 3.15: Realistic texture replacement: a virtual painting (left) is added, and a brick texture (right) applied. The textures realistically interact with the illumination (red circles). With a naïve texturing approach, shadows are lost.

3.7.2 Editing Material Appearances

This application demonstrates the modification of material properties other than reflectance at real-time rates; The term material editing is borrowed from Ye *et al.* (2014), who showed similar effects in an offline setup. Tone mapping is applied to a selected region of the shading layer that has been computed in real-time. The tone mapping function is provided interactively by the user based on a sparse set of control points. Based on this, the appearance of different objects can be changed in live video footage (see Figure 3.14). The cushion is modified to have a velvet surface, whereas in the second image, the cloth is made to appear more shiny. Note, the reflectance of the objects is not influenced by this operation, since the editing is performed in the shading domain.

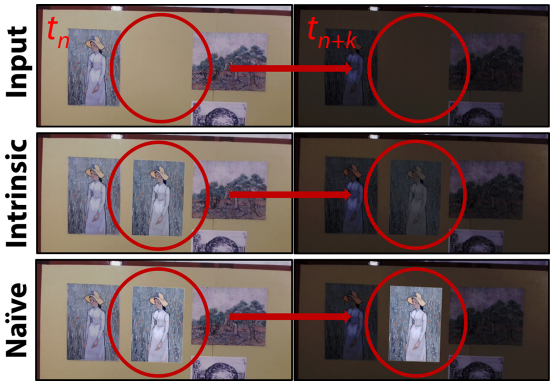


Figure 3.16: Realistic texture replacement: a virtual painting is added to the wall. The textures realistically reflect the illumination change (red circles) caused by dimming the lights. Note that a naïve texturing approach leads to unrealistic results.

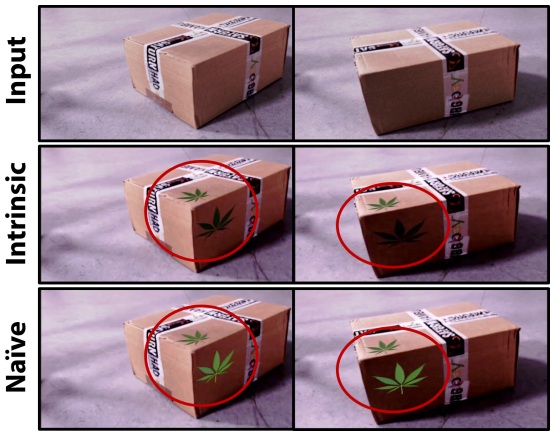


Figure 3.17: Realistic texture replacement: two virtual decals are added to a box. Intrinsic texturing realistically interacts with the real world shading. Note that naïve texturing leads to unrealistic results.

3.7.3 Realistic Texture Replacement

Real-time illumination-aware retexturing of live video footage is demonstrated. In contrast to the two previous examples, which applied a constant colour or appearance change to a chroma-keyed region, this demo requires temporal correspondences. To this end, the feature-based PTAM (Klein & Murray, 2007) technique which tracks the camera's motion based on a set of sparse feature points in real-time is used. Retexturing is performed by applying a reflectance layer texture to the handled planar surfaces. Note however, arbitrary objects can be handled easily if a corresponding proxy geometry is available. In Figure 3.15, a Van Gogh painting ("Girl in White, 1890") is added to the scene. The intrinsic retexturing method adds shadows and lighting, which are part of the real scene, to the texture in real-time. This allows for photorealisitic results. The naïvely added texture, i.e. replacing the texture in the non-decomposed RGB video, does not interact with the illumination, hence making it appear synthetic. The notice board is also retextured with a brick texture. In Figure 3.16, the light source is dimmed. This approach properly relights the synthetic

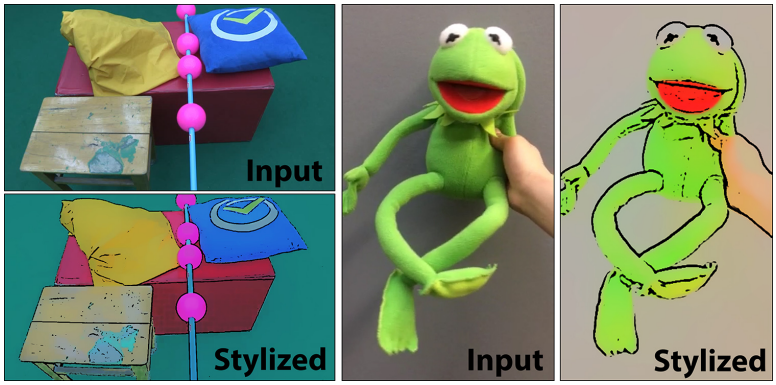


Figure 3.18: Live video stylization using a cartoon-style effect.

texture. Note, the virtual paintings and bricks are correctly and realistically interacting with the real world illumination. In contrast, naïve retexturing leads to unrealistic results. In Figure 3.17, a leaf texture is added to the side of a carton. Note the different shading on the added decal, depending on which side of the box it is placed.

3.7.4 Live Video Abstraction & Stylization

Next, an abstraction and artistic stylization of live video footage is demonstrated. Abstraction of images and video has been shown to be an important tool in recognition and memory tasks (Winnemller *et al.*, 2006). The reflectance video stream does not contain shading information and hence already captures an abstract version of the scene. By increasing the contrast of major edges of the shading layer, and suppressing low-contrast regions, a nice cartoon-style effect can be achieved. To this end, a difference-of-gradient (DoG) filter (Winnemller *et al.*, 2006) is applied to the shading layer and then recombined with the reflectance layer (see Figure 3.18). The spatial scale, sensitivity and sharpness of the resulting edges can all be controlled interactively by the user. Unlike previous video abstraction techniques, this method is directly applied to the shading layer, hence enforcing only the shading edges, not edges between albedo regions which are often also stylized in previous methods.

3.8 Discussion

The first approach for intrinsic decomposition of live video streams at real-time frame rates was demonstrated. While high-quality results on par or surpassing the current state-of-the-art offline methods in terms of robustness, accuracy and

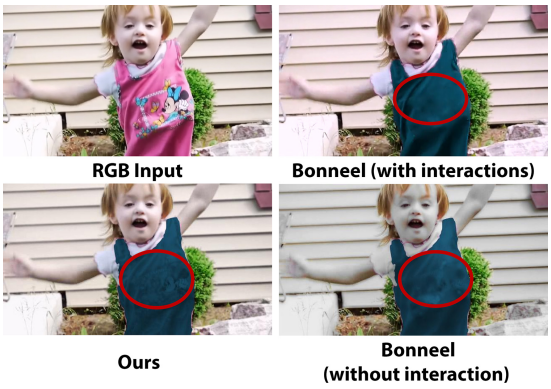


Figure 3.19: Recolouring of highly textured objects: comparable recolouring results to the approach of Bonneel *et al.* (2014) with default parameters and no user-interaction (bottom) are obtained. With additional scribble-based user-interaction, Bonneel *et al.* obtain results with fewer texture copy artifacts (top right). Note that this approach is one order of magnitude faster and does not use any user input, since this is infeasible in the proposed live video editing context.

runtime are achieved, some simplifying assumptions are made to make this hard inverse problem tractable. Note that these assumptions are common to almost all state-of-the-art intrinsic decomposition approaches, even to the offline methods. In the following, the main assumptions are discussed:

Monochromatic Illumination: All illuminants are assumed to emit pure white light, a reasonable assumption for many real world scenes. Therefore, a perceived change in chromaticity can be directly attributed to a change in material reflectance.

Diffuse Reflectance: All objects in the scene are assumed to have a purely diffuse reflectance. This is a soft assumption since the method handles non-diffuse objects gracefully, as long as the material is not highly specular.

Sparse Reflectance: The scene is assumed to be comprised of a relatively small number of uniformly coloured surface patches. In natural scenes with high-frequency texture or smooth colour gradients, this assumption might be violated. One such example in the context of recolouring is shown in Figure 3.19.

Direct Illumination: Only direct illumination effects are considered. Complex multi-bounce illumination such as caustics or colour bleeding (which will be tackled in Chapter 5) are not explicitly handled and might be mistaken for reflectance variation.

Despite these simplifying assumptions, this approach produces plausible decomposition results at previously unseen frame rates and without any user-interaction.

3.9 Conclusion

This chapter introduces the first approach to compute intrinsic decompositions of monocular live video footage in real-time. High-quality and temporally coherent decompositions are obtained without the need for an explicit correspondence search. Real-time optimization is possible due to a carefully crafted data-parallel solver for general $\ell_2\ell_p$-optimization problems. The capabilities of the approach are demonstrated on live video footage as well as on synthetic data. The qualitative and quantitative evaluation shows that the approach is on par with or even outperforms state-of-the-art techniques in terms of robustness, accuracy and runtime. The real-time capabilities of this intrinsic decomposition approach can pave the way for many novel augmented reality applications that build on top of the presented realistic recolouring, relighting and texture editing functionality.

The next chapter builds on this work by incorporating depth information from a depth sensor to combine intrinsic decomposition with geometry reconstruction. This allows for direct user-interaction with the scene to improve the decomposition at very ambiguous locations and extends the applications shown in this chapter to 3D objects, enabling new real-time applications such as relighting. Subsequent chapters of this thesis aim to relax the assumptions to make the approach applicable to an even wider range of settings, such as multi-bounce illumination and highly specular surfaces.

Chapter 4

Live User-Guided Intrinsic Video

In Chapter 3, it was shown that real-time intrinsic decomposition of video streams allows for several new augmented reality applications. But without the knowledge of scene geometry, the method is inherently limited in its abilities. Additionally estimating scene geometry could lead to a solution to the full rendering equation. This chapter presents the first method to simultaneously solve for intrinsic decomposition and the scene geometry of a static scene in unconstrained settings using a single RGBD sensor (Meka *et al.*, 2017a). This further enhances real-time scene editing capabilities such as direct user-interaction in 3D that can be propagated across the scene geometry and applications such as real-time photorealistic relighting.

4.1 Introduction

The ability to edit the appearance of the real world seen through a mobile device or a head-mounted see-through display – such as photorealistic recolouring and relighting of real scenes – is essential for many augmented reality (AR) applications. Imagine a virtual refurnishing application that allows a user to roam around and explore different colour choices for real world objects, or different placements of virtual lights, directly in their living room.

To enable this, an AR system needs to jointly track its position and reconstruct the geometry of the scene – initial solutions to this hard problem exist. The much harder problem, however, is that the system needs to solve a complex inverse rendering problem in real-time. Ideally, from monocular or RGB-D sensors alone, the AR device has to estimate detailed models of surface reflectance and scene illumination, in order to modify both through computer graphics overlays. As of

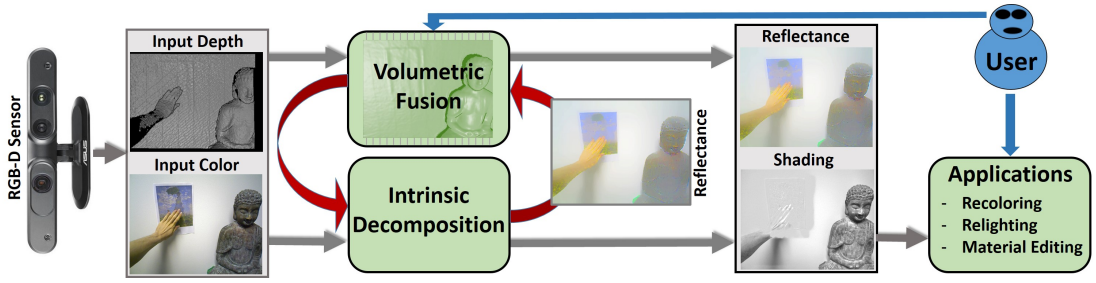


Figure 4.1: This novel user-guided intrinsic video approach enables real-time applications such as recolouring, relighting and material editing.

today, estimating *fine-grained light transport models* in *general unconstrained scenes* from *only one on-board camera* is far from possible in *real-time*.

The problem of intrinsic video decomposition was tackled for the first time by Bonneel *et al.* (2014); Ye *et al.* (2014), but at slow offline frame rates. In Chapter 3 it was shown that this task can be solved in real-time. However, intrinsic decomposition is ill-posed (Barrow & Tenenbaum, 1978b), as the separation of reflectance colour and illuminant colour is ambiguous. On heavily textured objects, this inherent ambiguity leads to reflectance texture information being erroneously 'baked' into the shading. In addition, high-frequency shading effects are often misinterpreted as texture. Although regularization of reflectance and shading can reduce such artifacts, they cannot be entirely resolved in the existing techniques. This leads to visual artifacts in the targeted AR applications.

To alleviate this problem and enable live realistic editing of reflectance and lighting in augmented reality, a novel interactive scene-level approach is proposed here for real-time intrinsic decomposition of static scenes captured by an RGB-D sensor. This approach is based on a volumetric representation of the scene that serves as a proxy to store reflectance estimates and sparse user-provided constraints, such as constant shading and constant reflectance strokes, directly in 3D. This has several fundamental advantages compared to 2D-based approaches. Since the user constraints are stored in 3D space, they can be robustly re-projected to arbitrary novel views of the scene to further constrain the intrinsic decomposition problem. Similarly, surface reflectance estimates are densely fused into the volumetric reconstruction of the scene. This enables the re-projection of the estimates to novel views to further constrain and jump-start the intrinsic decomposition process. This also leads to more temporally stabilized decomposition results as demonstrated in Section 4.7. Another benefit of fusing reflectance estimates is that complete coloured 3D models that are devoid of shading information can be obtained.

The user constraints are intuitively provided directly on the real world geometry with a touch-based interaction metaphor that allows to apply strokes on the 3D geometry or via live mouse interactions on the 2D image. By providing shading and reflectance strokes, the user can interactively improve the decomposition result to resolve ambiguities and obtain higher quality results than with previous fully-automatic approaches.

The obtained decomposition quality improves on quality shown in Chapter 3, which does not consider user input and thus suffers particularly in highly textured regions. Since this approach runs live, the user can reexamine the decomposition result at any time, and place additional strokes if required. In addition to photorealistic recolouring and material editing, the availability of the underlying geometry model, which is jointly recovered, enables advanced editing effects, such as physically correct relighting of objects. Note that the proposed approach can also be used to decompose standard RGB images, without geometry, in real-time, since it can also directly use 2D mouse strokes as constraints. The motivation of this work is not interaction design itself, but the design of the algorithms that enable the use of scene geometry and user-interaction for enhanced intrinsic decomposition of a live video stream as well as its photorealistic augmentation. In summary, this method is based on the following main technical contributions:

- A volumetric scene representation to densely store the obtained reflectance estimates, and user-provided strokes for constant shading and reflectance in 3D.
- A real-time intrinsic decomposition approach that exploits these constraints to solve the ill-posed decomposition problem.
- Prototype augmented reality applications such as live recolouring, material editing and relighting at high-quality.

4.2 Related Work

Intrinsic image decomposition has a long history, stretching from the seminal Retinex approach of Land & McCann (1971) all the way to the current day. The key insight of Retinex, which helps to disambiguate shading and reflectance, is that larger image gradients mostly correspond to gradients in reflectance rather than shading. Therefore, thresholding the image gradients can be used for disambiguating these two components. This idea has been extended by using learned classifiers instead of fixed thresholds by Tappen *et al.* (2005), it has also been combined with non-local cues for improving decompositions by Gehler *et al.* (2011); Shen & Yeo (2011), and

closed-form solutions have been proposed by Zhao *et al.* (2012). To further improve the quality of intrinsic decompositions, additional, increasingly complex priors have been proposed to constrain the solution space. Many techniques assume that the reflectance distribution in a scene is sparse (Ding *et al.*, 2017; Gehler *et al.*, 2011; Shen & Yeo, 2011; Shen *et al.*, 2013), i.e. that there are only a few different colours visible at the same time, which can for example be determined using clustering of Garces *et al.* (2012), efficient inference via dense conditional random fields such as in Bell *et al.* (2014) or image flattening as in Bi *et al.* (2015). Barron & Malik (2015a) even model and estimate shape and illumination in addition to reflectance, and Kong & Black (2015) also estimate contours, depth and optical flow from videos. Recently, more advanced reflectance priors have been learned directly from ground-truth intrinsic decompositions by Zhou *et al.* (2015); Zoran *et al.* (2015). Image sequences can also provide temporal constraints, for example when reflectance is constant but shading varies temporally, such as in Kong *et al.* (2014); Laffont & Bazin (2015); Matsushita *et al.* (2004); Weiss (2001). The additional depth channel captured by consumer depth cameras has also been exploited to provide additional constraints by Chen & Koltun (2013); Hachama *et al.* (2015); Lee *et al.* (2012). The proposed approach also uses a depth camera for enabling scene-consistent temporal propagation of reflectance estimates and user constraints that are densely stored in the reconstructed scene geometry.

In many cases, the existing priors fail to achieve intrinsic decompositions of high-quality. Annotations such as scribbles provided by a user can help to guide the intrinsic decomposition towards the desired solution, as shown by Bousseau *et al.* (2009); Shen *et al.* (2011). Previous approaches for offline intrinsic video decomposition also make use of scribbles to obtain higher quality decomposition results. Ye *et al.* (2014) use a scribble-based technique for decomposing the first video frame, and Bonneel *et al.* (2014) allow strokes for any video frame and use them as necessary. In contrast, the approach shown previously in this thesis in Chapter 3 explicitly excludes scribbles as they cannot be provided in real-time at 30 Hz. In this work, it is shown how to make scribbles work in a live setup by embedding them in a dense volumetric 3D reconstruction of the scene that makes them independent of the current camera viewpoint. In addition, the reconstruction is used as a proxy to fuse and temporally propagate reflectance estimates.

Placing virtual objects into a real world scene in a seamless, photorealistic fashion requires an accurate estimate of the incident scene illumination, so that the virtual object can be lit plausibly. Gruber *et al.* (2012) use spherical harmonics from the colour observations of a jointly reconstructed 3D scene model, which enables

plausible illumination effects in a virtual-reality rendering pipeline. Follow-up work extended this approach to dynamic real world scenes within an augmented reality rendering pipeline (Gruber *et al.*, 2015).

Scene reconstruction based on commodity RGB-D sensors employs spatiotemporal filtering (Richardt *et al.*, 2012) or implicit surface representations (Curless & Levoy, 1996; Fuhrmann & Goesele, 2014; Zhou & Koltun, 2013), since they allow to deal with the noisy depth data captured by commodity sensors. The first on-line method for the reconstruction of a static scene using a hand-held depth sensor was KinectFusion by Izadi *et al.* (2011); Newcombe *et al.* (2011). The scene's geometry is approximated using a truncated signed distance field (Curless & Levoy, 1996), into which per-frame data is fused. Camera tracking is based on a fast variant of the iterative closest point (ICP) algorithm by Rusinkiewicz & Levoy (2001) that uses projective correspondences. Many approaches that extend KinectFusion have been proposed, with the focus on extending the scale of the limited reconstruction volume such as Chen *et al.* (2013); Newcombe *et al.* (2011); Roth & Vona (2012); Steinbrcker *et al.* (2014); Zeng *et al.* (2013). Another extension performs an illumination-invariant 3D reconstruction (Kerl *et al.*, 2014) using time-of-flight RGB-D cameras: illumination-independent surface reflectance is first computed in the infrared domain, and then transferred to the colour domain by solving an optimization problem. This approach enables the placement of user constraints via interactions in real-time, which allows to incrementally improve the decomposition results, and alleviates the texture copy problem. The SemanticPaint approach of Valentin *et al.* (2015) combines dense volumetric reconstruction with a learning-based segmentation strategy to obtain a semantic decomposition of the scanned scene.

4.3 Overview

A high-level overview of this novel user-guided intrinsic video approach is shown in Figure 4.1. First, a volumetric representation of the scene is reconstructed using the RGB-D data captured by a commodity depth sensor. To this end, the dense volumetric 3D reconstruction approach of Niener *et al.* (2013) is employed that obtains high-quality reconstructions of static scenes in real-time (Section 4.4). In contrast to Niener *et al.* (2013), here, this representation is used as a proxy to densely fuse surface reflectance estimates instead of the input image colours. For this, an intrinsic decomposition of the colour video stream is simultaneously computed during volumetric reconstruction, and the obtained reflectance estimates fused into the

reconstruction. The fused reflectance information is used to further inform the underlying intrinsic video decomposition problem. In addition to the surface reflectance, user-provided constraints are also stored in the form of constant reflectance and shading strokes. Such constraints can be provided using live mouse input or using an intuitive touch-based interaction metaphor. In the case of touch-based input, the user is automatically detected by a foreground segmentation approach that utilizes the difference in geometry and colour between the reconstructed model and the current input RGB-D frame. This allows us to detect touch-based user-interaction on real world geometry, and enables the user to interactively place constraints in the scene (Section 4.5). These constraints are projected into novel views to further constrain the ill-posed intrinsic decomposition problem (Section 4.6). The proposed approach facilitates a variety of augmented reality applications, such as recolouring, material editing and relighting (Section 4.8).

4.4 Representation

As the user walks around a scene with an RGB-D camera, which could for example be integrated into a head-mounted AR device, a virtual model of the scene is obtained using VoxelHashing, a large-scale dense volumetric reconstruction technique (Niener *et al.*, 2013). The source code for the VoxelHashing framework is publicly available[1]. The captured depth maps are fused into a high-quality model using a truncated signed distance field (Curless & Levoy, 1996) (4 bytes per voxel). Memory is managed based on a space and time efficient spatial hashing strategy. Internally, 3D space is discretized into a set of discrete voxels, which are stored as blocks consisting of $8^3 = 512$ voxels each. The camera's rigid motion is tracked using a fast variant of the iterative closest point (ICP) algorithm that uses projective correspondence lookups.

In contrast to Niener *et al.* (2013), the observed colour samples are not fused in the volume, but directly fuse surface reflectance estimates (12 bytes per voxel). To this end, the simultaneously captured colour image is decomposed into its shading and reflectance layers (Section 4.6). The surface reflectance is devoid of illumination information and is fused using temporal exponential averaging. Since multiple per-frame reflectance estimates are averaged, sensor noise and inconsistencies in the decomposition results are significantly reduced. In addition to surface reflectance, the dense volumetric reconstruction is used to store additional user-provided constraints based on a stroke identifier (1 byte per voxel). Storing the

[1] https://github.com/niessner/VoxelHashing

constraints directly in 3D world space allows us to re-project them to arbitrary novel camera views, and hence help to solve the ill-posed intrinsic decomposition problem. In addition, the spatial neighbourhood information is encoded in the volumetric grid to propagate constraints in 3D space (see Section 4.5.2) to obtain a basic segmentation of the scene. This is useful for applying constraints directly to larger parts of the scene, and is used in several proposed AR applications (see Section 4.8).

4.5 Interactions

After reconstruction of the scene's geometry, the user can interact with the scene using live mouse input or a touch-based interaction metaphor to provide constraints to further inform the ill-posed intrinsic decomposition problem. Constraints are given in the form of constant reflectance and constant shading strokes. The dense volumetric reconstruction of the scene is used as a proxy to store the constraints directly in world space on a per-voxel level (using a stroke identifier attribute). For example, the user can place a constant shading stroke on a wall to alleviate the texture copy problem encountered in previous approaches, where high-frequency reflectance is often erroneously copied to the shading layer. In addition to these constraints, the user input is used in the proposed live AR applications (see Section 4.8), where it enables recolouring, material editing and relighting. The user can for example simply touch a chair to assign a different colour to it, or change the material of any object in the scene. In summary, all supported interactions are:

- **Constant Shading Stroke:** All surface points belonging to the same stroke are enforced to share the same shading value. Multiple independent strokes of this type can be used.
- **Constant Reflectance Stroke:** This constraint enforces all associated surface points to share the same reflectance colour. Multiple independent strokes of this type can be defined.
- **Recolouring Stroke:** The reflectance of all associated surface points is set to a fixed user-specified colour. Using this stroke type, users can paint an arbitrary reflectance map.

All strokes optionally support a region filling strategy that allows to directly select a complete subset of the scene. The propagation of stroke attributes is based on spatial connectivity and reflectance similarity, as detailed below.

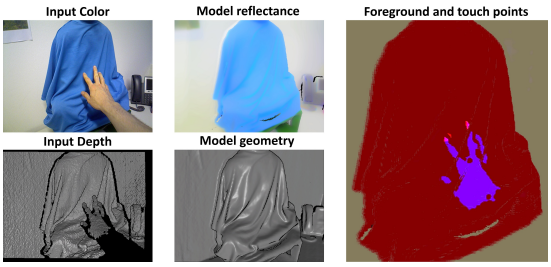


Figure 4.2: The user's hand is detected as foreground based on the difference between the input depth image and the reconstructed scene. Touch points are detected (bright red) and propagated based on spatial connectivity and reflectance similarity.

4.5.1 Detection of Touch Points

Once scene reconstruction is finished, the integration of further geometry is stopped to allow the user to interact with the obtained reconstruction of the scene by placing constraints. Interactions are based either on live mouse input or a touch-based interaction metaphor. Touch-based interaction requires the user to closely interact with objects that are in plain view of the camera (see Figure 4.2). This might throw off the camera tracker, since the motion of the user violates the assumption that the scene is static. To alleviate this problem, the pixels that correspond to the user are automatically detected and exclude them from tracking. To this end, correspondences in the ICP alignment strategy are pruned if the distance between points is larger than $\epsilon_{\text{dist}} = 15\,\text{cm}$ or the normals deviate more than $\epsilon_{\text{norm}} = 14$ degrees. After alignment, all outliers in the input depth map are considered foreground. In the next step, touch points are determined based on the spatial proximity of the background and skin-coloured foreground pixels similar to Vezhnevets *et al.* (2003). For every detected touch point, all voxels that fall within a small spherical neighbourhood are marked, similar to a 3D brush. The radius of the sphere can be controlled by the user. In the case of live mouse input, the rendered depth map is used to back-project the strokes to 3D space.

4.5.2 Spatial Constraint Propagation

To enable fast and convenient editing, the user is provided the option to automatically propagate constraints to appropriate spatial subregions. Semantic segmentation is a challenging and ill-posed problem, especially at real-time rates. The SemanticPaint approach of Valentin *et al.* (2015) presents an impressive solution to this problem, but has a high memory footprint and is already quite computationally demanding. Since the goal of this approach is user-guided intrinsic decomposition, a lightweight segmentation approach is used, which leaves enough computational resources for the other processing steps. For each stroke, the average reflectance value of all influenced voxels is computed and stored for further processing. A data-parallel

flood fill is performed to all neighbouring voxels that have a sufficiently similar reflectance in RGB colour space to the stored value.

The data-parallel flood fill works by propagating a 3D voxel frontier starting from a seed point. The current frontier is managed using an array in global device memory that implements a list of voxels. The insertion of new elements into the list is managed using an atomic counter. Given the current frontier, the frontier is advanced in space by starting one thread per voxel in the list. Each thread examines its $3 \times 3 \times 3$-voxel neighbourhood. A binary mask is used to store which voxels have already been processed, and append unprocessed neighbouring voxels that fulfill the flood fill criterion to a new voxel frontier list. In the flood fill criterion, reflectance similarity is thresholded based on the distance in RGB colour space ($\epsilon_{\text{fill}} = 0.1$) between the reflectance of the currently processed voxel and the stored average reflectance of the current stroke. Note that the flood fill implicitly takes the connectivity of the sparse voxel grid into account.

4.6 Energy

The majority of intrinsic video decomposition techniques suffer from the texture copy problem, leading to residual texture in the shading layer. This is because it is highly challenging to correctly disambiguate texture into its reflectance and shading components in the absence of additional constraints. A number of intrinsic decomposition techniques such as Bonneel *et al.* (2014); Bousseau *et al.* (2009); Shen *et al.* (2011); Ye *et al.* (2014) have therefore resorted to user-interaction in the form of strokes to provide additional constraints. It is proposed to use live mouse interactions or a touch-based interaction metaphor directly in 3D space for intuitive editing of the intrinsic decomposition. User input is stored densely based on the obtained scene reconstruction. In addition, computed reflectance estimates are fused using the volume. At run time, reflectance estimates and constraints are projected to novel views to constrain the ambiguous intrinsic decomposition problem towards a higher quality solution. Previous constraint-based approaches run offline and require long computation times. In contrast, the proposed approach runs at real-time frame rates, thus making it usable in the augmented reality context.

4.6.1 Variational Intrinsic Video Decomposition

Finding the optimal intrinsic decomposition $\mathbf{D}^*$ is cast as a general non-linear energy minimization problem:

$$\mathbf{D}^* = \underset{\mathbf{D}}{\operatorname{argmin}} E(\mathbf{D}) \tag{4.1}$$

$$\mathbf{D} = \left[...,\mathbf{r}(\mathbf{x})^\top,...,s(\mathbf{x}),...\right]^\top, \tag{4.2}$$

where the vector $\mathbf{D}$ contains all unknowns, i.e. log-space reflectance $\mathbf{r}(\mathbf{x}) \in \mathbb{R}^3$ and shading $s(\mathbf{x}) \in \mathbb{R}$ for all pixels $\mathbf{x} \in \Omega \subset \mathbb{N}^2$. The employed intrinsic video decomposition energy E is based on several objective functions:

$$E(\mathbf{D}) = E_{\text{fit}}(\mathbf{D}) + w_{\text{r}} E_{\text{reg}}(\mathbf{D}) + w_{\text{u}} E_{\text{user}}(\mathbf{D}) + w_{\text{s}} E_{\text{stab}}(\mathbf{D}). \tag{4.3}$$

The objective functions model the reproduction of the input image E_{fit}, spatio-temporal regularization E_{reg}, integration of the user constraints E_{user}, and temporal stabilization E_{stab}. The constant weights $w_{\text{r}} = 1$, $w_{\text{u}} = 1000$ and $w_{\text{s}} = 10$ control the influence of the different objectives. In the following, all terms are discussed in more detail.

Reproduction of the Input Image The fitting objective E_{fit} enforces that the decomposition reproduces all N pixels of the input colour image $\mathbf{I}$. This constraint is formulated in the log-domain for linearity:

$$E_{\text{fit}}(\mathbf{D}) = \sum_{\mathbf{x} \in \Omega} \left\| \mathbf{i}(\mathbf{x}) - \left(\mathbf{r}(\mathbf{x}) + [1\ 1\ 1]^\top s(\mathbf{x}) \right) \right\|^2. \tag{4.4}$$

Here, $\mathbf{i}(\mathbf{x}) = \ln \mathbf{I}(\mathbf{x}) \in \mathbb{R}^3$ is the logarithm of the pixel colour at pixel $\mathbf{x}$, $\mathbf{r}(\mathbf{x}) = \ln \mathbf{R}(\mathbf{x})$ is the log-reflectance, and $\mathbf{s}(\mathbf{x}) = \ln \mathbf{S}(\mathbf{x})$ the log-shading value of the same pixel.

Spatio-Temporal Regularization For regularization, the spatio-temporal strategy discussed in the previous chapter (Section 3.4.3) is followed and a combination of four different terms is employed:

$$E_{\text{reg}}(\mathbf{D}) = w_{\text{p}} E_{\text{p}}(\mathbf{D}) + w_{\text{g}} E_{\text{g}}(\mathbf{D}) + w_{\text{m}} E_{\text{m}}(\mathbf{D}) + w_{\text{c}} E_{\text{c}}(\mathbf{D}). \tag{4.5}$$

Since many man-made scenes contain a small, distinct number of reflectance values, sparsity based on a p-norm constraint is enforced:

$$E_{\mathrm{p}}(\mathbf{D}) = \sum_{\mathbf{y}\in\mathbf{N}(\mathbf{x})} \omega_{\mathrm{cs}}(\mathbf{x},\mathbf{y}) \cdot \|\mathbf{r}(\mathbf{x}) - \mathbf{r}(\mathbf{y})\|_2^p, \tag{4.6}$$

where $\omega_{\mathrm{cs}}(\mathbf{x},\mathbf{y}) = \exp(-15 \cdot \|\mathbf{c}(\mathbf{x}) - \mathbf{c}(\mathbf{y}))\|_2)$ measures the chroma similarity of two adjacent pixels. Spatio-temporal coherence is incorporated based on a global prior E_{g} that takes long-range chroma similarity into account. The prior E_{m} enforces ℓ_2-spatial smoothness of the shading layer at chroma boundaries. Finally, a soft constraint on chroma similarity E_{c} keeps the chroma of the reflectance image close to the chroma of the input. For a detailed discussion of the terms $E_\bullet$, the sparsity norm ℓ_p and parameters $\omega_\bullet$, refer to the corresponding section in the previous chapter Section 3.4.2.

Incorporation of User Constraints One of the main contributions of this work is a novel approach for incorporating the user constraints, in the form of constant reflectance and constant shading strokes, directly into the optimization problem:

$$\begin{aligned} E_{\mathrm{user}}(\mathbf{D}) = &\sum_{\mathbf{S}_i \in S} \sum_{\mathbf{x}\in\mathbf{S}_i} w_i(\mathbf{x}) \cdot |s(\mathbf{x}) - \hat{s}_i|^2 + \\ &\sum_{\mathbf{R}_i \in R} \sum_{\mathbf{x}\in\mathbf{R}_i} w_i(\mathbf{x}) \cdot \|\mathbf{r}(\mathbf{x}) - \hat{\mathbf{r}}_i\|^2. \end{aligned} \tag{4.7}$$

Here, S is the set of shading strokes, and $\mathbf{S}_i$ the set of pixels belonging to the i-th shading stroke. $\hat{s}_i$ is the representative unknown shading value associated with the i-th stroke. The same notation holds for the reflectance strokes $\mathbf{R}_i \in R$. Note that $\hat{s}_i$ and $\hat{\mathbf{r}}_i$ are unknown auxiliary variables. For the i-th stroke, a per-pixel stroke weight $w_i(\mathbf{x})$ is defined to fade out the influence of the strokes (squared fall-off) close to their boundary.

Stabilization of Reflectance Estimates Another important contribution of this work is to densely fuse the obtained reflectance estimates into the volumetric scene representation. This allows to enforce temporal coherence and further inform the intrinsic decomposition problem. To this end, the following novel stabilization constraint is proposed:

$$E_{\mathrm{stab}}(\mathbf{D}) = \sum_{\mathbf{x}\in\Omega} B(\mathbf{x}) \cdot \left(\mathbf{r}(\mathbf{x}) - \mathbf{r}^{\mathrm{model}}(\mathbf{x})\right)^2. \tag{4.8}$$

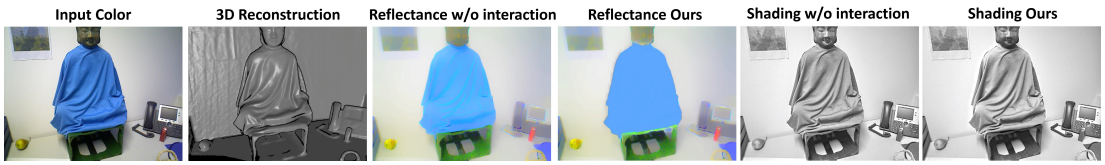


Figure 4.3: Constant reflectance strokes improve the decomposition by moving the high-frequency shading of the cloth to the shading layer.

The background mask $B(\mathbf{x})$ (one for background, zero for foreground) prunes any potentially dynamic foreground pixels. It encourages the per-pixel log-reflectance values $\mathbf{r}(\mathbf{x})$ to be close to the fused mean log-reflectance $\mathbf{r}^{\text{model}}(\mathbf{x})$ stored in the volumetric scene model. This per-pixel mean log-reflectance is computed by extracting the log-reflectance-coloured isosurface of the volumetric scene representation via ray marching.

4.6.2 Data-Parallel Optimization

The goal is to compute the intrinsic decomposition of the RGB-D video stream at real-time frame rates. Therefore a fast and efficient strategy is required to solve the underlying non-linear optimization problem. A highly efficient, data-parallel, iteratively reweighted least squares (IRLS) solver is proposed that allows computing the optimum of the energy E (see Equation (4.3)) at real-time rates. In contrast to the approach described in the previous chapter (Chapter 3), this decomposition objective does not have a sparse Jacobian, but contains dense blocks due to the incorporation of the user constraints E_{user} (see Equation (4.7)). This is because every per-pixel unknown belonging to the same stroke has a derivative with respect to the unknown per-stroke auxiliary variable. Therefore, the data-parallel solver proposed in previous work is not sufficient to achieve real-time frame rates, since the workload is not equally distributed between different threads. Such a joint optimization of all variables would lead to faster convergence with respect to the required number of iteration steps, but every iteration of the solver would require significantly more time, since the data-parallel compute power of the GPU can not be fully exploited.

To tackle this problem, an iterative flip-flop strategy is proposed that solves two simpler optimization problems in alternation. Given an initial estimate of the per-pixel shading and reflectance, the auxiliary variables are first optimized. Since the auxiliaries only appear in the least-squares objective E_{user}, the optimum has a closed-form solution, and can be obtained as the average of all associated per-pixel values that belong to the same stroke. After this, the new values are fixed for the auxiliaries,

and a new decomposition is optimized using a data-parallel IRLS solution strategy. As the auxiliaries are constant during this step, the Jacobian is again sparse, leading to high performance. Convergence is assumed after 7 iteration steps. Internally, the IRLS solver divides the problem into small rectangular sub-domains and uses a data-parallel variant of the alternating Schwarz procedure similar to Zollhfer *et al.* (2015). Each iteration solves the local problems in shared memory and exchanges data with neighbouring domains using global memory. This optimization is applied using a coarse-to-fine strategy (5 levels) for faster convergence. Starting from the coarsest level, the coarse scale version of the problem is solved to obtain an approximate solution. This solution is then upsampled to the next finer level and use for initialization.

4.7 Experiments

The proposed approach is demonstrated in a live setup. A PrimeSense Carmine 1.09 close-range RGB-D sensor is used to obtain two 640×480 video streams of colour and depth at 30 Hz. Note that this approach is agnostic to the specific RGB-D sensor being used. Only a spatially and temporally aligned colour and depth stream is required as input. After acquiring an initial geometric model of the scene using dense volumetric reconstruction, the decomposition quality is improved using strokes that enforce constant reflectance and shading. The test scenes, including the results obtained by this approach, are publicly available on the project page[1] to encourage follow-up work and enable others to easily compare to this approach.

Decomposition Results Intrinsic decomposition of a scene allows for independent modification of the underlying physical layers of a scene while preserving the photorealism on reconstruction. Even for simple scenes with uniform, untextured regions, such photorealism cannot be obtained by simple luminancechrominance decomposition due the problem of 'chromaticity shift', as described in the previous chapter (Section 3.4.4). For textured surfaces, current state-of-the-art intrinsic decomposition approaches suffer from the texture-copy problem, if they do not rely on additional user input. Texture-copy refers to texture variation being misinterpreted as shading, as in Figure 4.4 (top). The proposed approach allows to resolve this problem via the incorporation of constant shading strokes into the decomposition problem, see Figure 4.4 (bottom). Without user input, it is difficult to disambiguate between blocks of varying intensity, and current state-of-the-art approaches fail in this regard. By adding

[1] http://gvv.mpi-inf.mpg.de/projects/InteractiveIntrinsicAR/

	Input	ICP	Decomposition	Fusion	Relighting
time (ms)	6.1	3.5	9.1	6.7	8.4

Table 4.1: Per-frame run time of the proposed user-guided intrinsic video approach averaged over the entire sequence for the scene in Figure Figure 4.7.

user constraints, the optimization approach better resolves the inherent ambiguities of the intrinsic decomposition problem, and a cleaner shading layer is obtained.

In addition, the decomposition of uniformly coloured regions suffers from the previously mentioned chromaticty-shift problem due to high-frequency shading variation, which is easily improved using a constant reflectance stroke, as can be seen in Figure 4.3. The clothing contains several dark creases that are wrongly contained in the reflectance layer in the absence of interaction. With an appropriate stroke, directly on the 3D geometry, this approach mitigates this issue and ensures constant reflectance over the cloth.

Runtime Performance and Memory Requirements For the reconstruction of the static scene geometry, a voxel resolution of 1 cm is used. Camera tracking takes 2 ms and reflectance fusion 0.6 ms. To project the user constraints into the image, ray marching is used, which takes 14 ms to compute the stroke map. Overall, the scene-level intrinsic decomposition runs at real-time frame rates. 5 hierarchy levels are used with 7 IRLS iterations on each. Each non-linear IRLS iteration performs 7 PCG steps internally. After each non-linear iteration, a flip-flop step is performed to update the auxiliary variables (Section 4.6.2). Intrinsic decomposition takes in total 22 ms per frame. While performing the 3D reconstruction of the scene, an average frame rate of 25 Hz is achieved. After the static reconstruction is completed, the frame rate increases to more than 30 Hz. All timings are computed on a commodity Nvidia GTX Titan graphics card. A more detailed breakdown of the timings of this approach is shown in Table 4.1.

This approach has a higher memory footprint than off-the-shelf VoxelHashing (Niener *et al.*, 2013), since the surface log-reflectance is stored using four bytes per colour channel. The memory footprint could be decreased by storing a discretized version of surface reflectance, e.g. one byte per colour channel. For the test scenes scenes, this is not the limiting factor, since more than 12 GB of global device memory are available on modern graphics hardware. For example, the sequence shown in Figure 4.7 requires in total 9.4 GB of global device memory for a voxel resolution

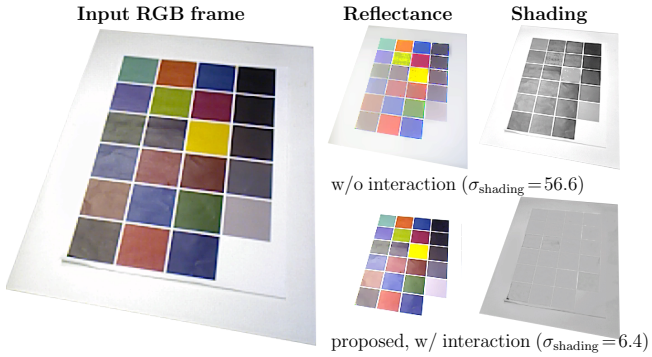


Figure 4.4: Intrinsic decomposition results for a colour chart. Without interaction, the shading image suffers from texture-copy. This approach improves the decomposition by using a constant shading stroke. This reduces the intensity variation of the shading layer (smaller standard deviation σ_{shading}).

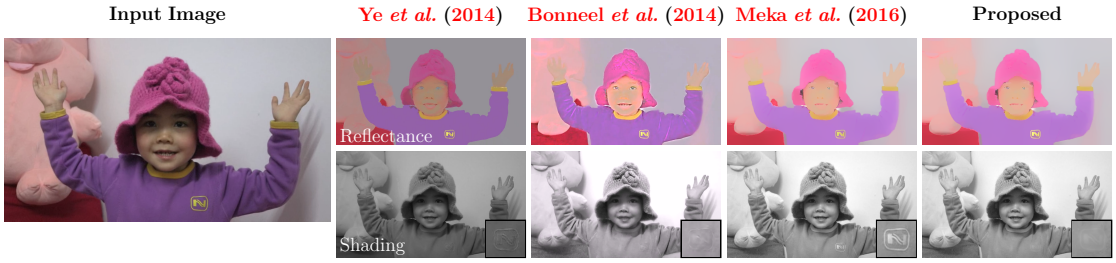


Figure 4.5: Comparison to state-of-the-art intrinsic video decomposition techniques on the 'girl' dataset. This approach matches the real-time performance of the approach shown in the previous chapter (Chapter 3, labelled in the figure as Meka *et al.* (2016)), while achieving the same quality as previous offline techniques of Bonneel *et al.* (2014); Ye *et al.* (2014) (see zooms).

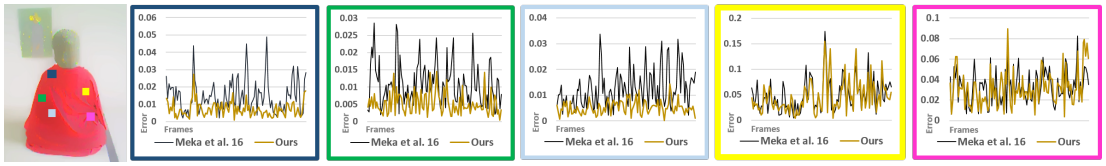


Figure 4.6: Temporal reflectance constancy: five rectangular regions are tracked and average albedo difference computed over time per region. The proposed approach uses fused reflectance estimates to further constrain and jump-start the intrinsic decomposition process. Therefore, it obtains a higher temporal reflectance consistency than the approach described in the Chapter 3 (labelled in the figure as Meka *et al.* (2016)).

of 4 mm (3.7 GB for a voxel resolution of 1 cm). This also includes all the data structures used during optimization and the memory to store the user constraints.

Reflectance Initialization In addition to the user constraints, surface reflectance estimates are also densely fused using the volumetric reconstruction of the scene. This allows to project the reflectance estimates to arbitrary novel views, which can be used to further constrain and jump-start the intrinsic decomposition process. Similar to the approach in the previous chapter, the reflectance layer is initialized in every

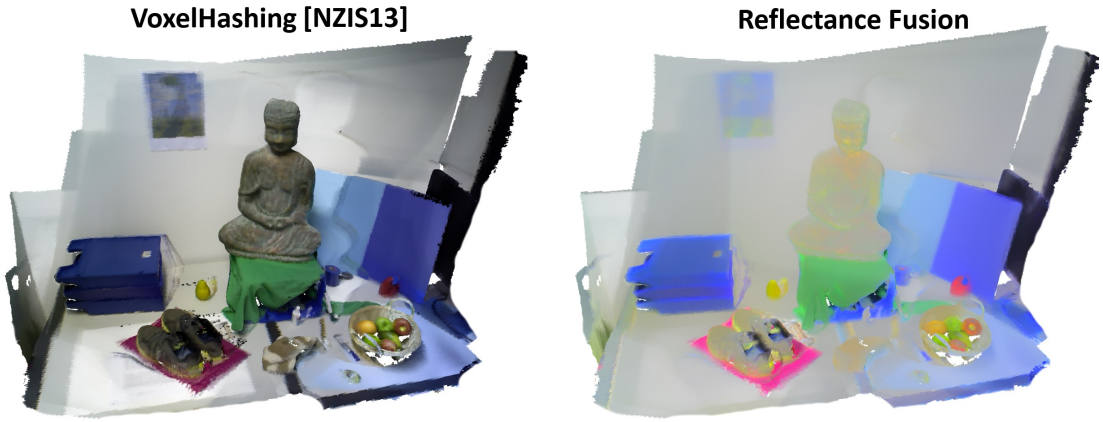


Figure 4.7: The proposed approach reconstructs the reflectance of the scene.

frame with the input RGB image, which could be far from the correct reflectance values. In contrast, this approach only uses this initialization for the first frame, and for subsequent frames synthesizes an initial reflectance map based on the projection of the fused reflectance estimates to the novel view. Occluded regions are initialized based on the corresponding input RGB values. The novel temporal stabilization term also helps to stabilize the decomposition results, see Figure 4.6. Five rectangular regions are tracked and the average albedo difference computed per region over time. As can be seen, this approach obtains a higher temporal reflectance stability than what was achieved without any novel-view reflectance initialization in Chapter 3 (average norm of albedo variation: 0.0187 instead of 0.0241). Another benefit of fusing reflectance estimates is that a complete coloured 3D model that is devoid of shading information is obtained, see Figure 4.7, which is in contrast to the colour reconstructed by state-of-the-art volumetric reconstruction technique of Niener *et al.* (2013).

Comparison to the State-of-the-Art The proposed approach is compared to the existing offline intrinsic video approaches by Ye *et al.* (2014) and Bonneel *et al.* (2014), which also use user-provided strokes for constraining the result, as well as the fully automatic real-time approach described in the previous chapter (Chapter 3). Note that these techniques work in a slightly less constrained setup, with a standard RGB camera, while this approach additionally leverages the available depth information of commodity RGB-D sensors. As these approaches operate on monocular colour video alone (without depth), the decomposition quality is compared on the 'girl' dataset in Figure 4.5, without using any geometry reconstruction and only 2D strokes within this approach. This comparison shows that the proposed approach

Figure 4.8: Photorealistic recolouring of a shirt.

obtains comparable decompositions to state-of-the-art offline approaches of Bonneel *et al.* (2014); Ye *et al.* (2014), but at real-time frame rates. The proposed decomposition quality improves on what was achieved in the previous chapter without considering user input, especially in regions with high texture variation, such as the logo on the shirt (see inset in Figure 4.5). Additional user constraints clearly help to resolve the inherent ambiguities of the intrinsic decomposition problem. Unlike existing methods, the proposed approach works best with RGB-D video streams, as strokes are placed directly on 3D geometry and can be projected to novel views of the scene for initializing them. Since this approach runs live, the user can reexamine the decomposition result at any time, and place additional strokes if required.

4.8 Interactive Applications

The proposed method enables a wide variety of interactive applications. In the following, several examples are shown, such as photorealistic recolouring, material editing and geometry-based relighting.

Photorealistic Recolouring and Material Editing Interactive and intuitive recolouring and material editing of real world objects is supported by this approach. Using the presented colour-based volumetric segmentation strategy, the complete geometry of the object that should be modified is first segmented. Since the segmentation is computed in 3D, the entire object can be segmented , even if it is not completely visible from the current view. The selected 3D geometry is projected to novel views to obtain the 2D mask that is later on used to modify the appearance of the object. For recolouring, an object's colour is replaced in the computed reflectance map of the current frame's decomposition by a user-defined colour. For material editing, a tone-mapping filter (as used by Ye *et al.* (2014)) is applied on the shading layer within the mask region. The modified layer is then recombined with the other intrinsic layer to

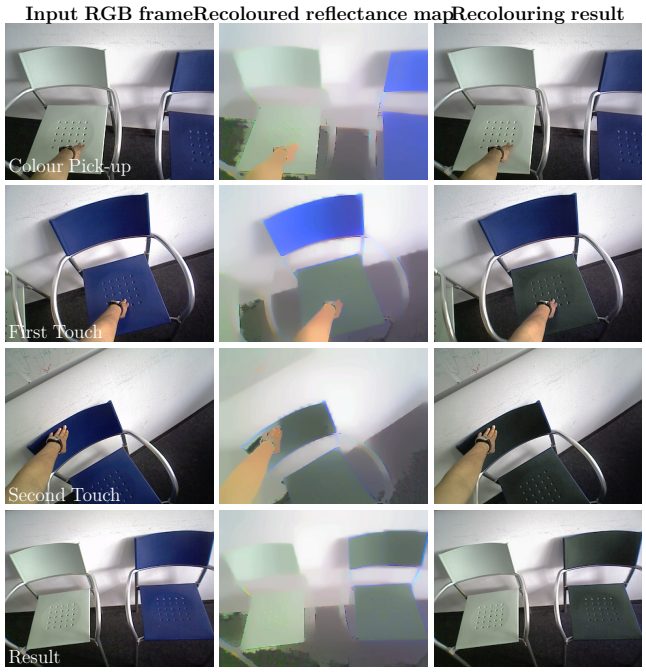


Figure 4.9: Interactive object recolouring. The reflectance of the light green chair is first picked up, and then transferred to the blue chair while preserving its brightness.

Figure 4.10: Dynamic geometry-based relighting. A virtual shading image is generated by rendering the scene geometry under a new light source. The resulting shading map is blended with the shading layer before recombining it with the reflectance to obtain a relighting effect.

obtain the final output, see Figure 4.8. By simply touching an object, the user can also choose to pick up a colour from the environment. This colour can then be used to recolour other objects, as illustrated in Figure 4.9. Instead of modifying the reflectance layer, a tone-mapping function can also be applied to the shading layer to change the appearance of an arbitrary object's material. This enables us for example to manipulate the appearance of a plaster cast such that it looks like metal, see Figure 4.11.

Geometry-Based Relighting In addition to modifications of the reflectance layer, geometry-based relighting is also presented, via modification of the shading

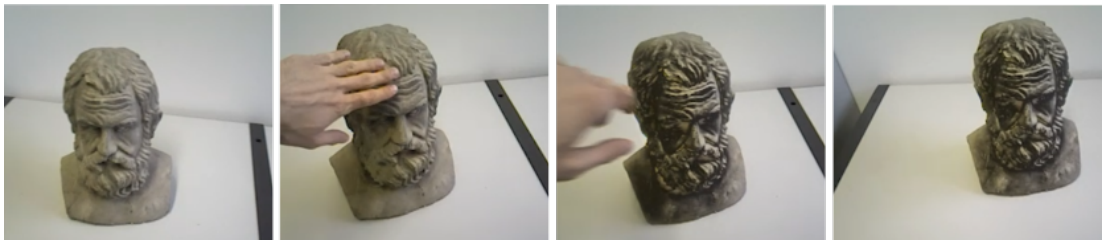


Figure 4.11: The shading layer is modified to convert plaster to metal.

layer. To this end, the user can place virtual light sources in the scene, which interact with it. The reconstructed 3D geometry is used in conjunction with the virtual light sources to generate a new shading image. The scene geometry is extracted using ray marching, and the synthetic shading map is computed by a fragment shader. The shading layer of this decomposition is blended with the synthesized shading map, then recombine the new shading layer with the reflectance map to obtain a compelling relighting effect, as shown in Figure 4.10.

4.9 Discussion

It is demonstrated that high-quality user-guided live intrinsic decomposition enables new scene modification applications. Still, the proposed approach has a few limitations. The geometric model of the scene is currently obtained beforehand in a pre-process, since it is required as the basis for foreground/background segmentation. In the future, alternative segmentation strategies can be developed.

This approach can only improve the decomposition quality of static scene geometry, since the user constraints are placed in 3D, and tracked based on a rigidly reconstructed scene model. Tracking dynamic time-dependent motion of non-rigidly deforming surfaces to also add and propagate constraints in such regions can be further investigated.

The improvement in decomposition quality via user constraints is of local nature, since the placed strokes only influence the decomposition result in a small surrounding neighbourhood. Therefore, similar to other stroke-based approaches, a lot of such constraints might be required to completely correct an initially very erroneous decomposition result. Fortunately, this is rarely the case and constraints are only required to deal with the highly textured regions of the scene.

The proposed simple touch-based interaction strategy sometimes leads to erroneous detections; more robust touch detection strategies are left for future work. Constraint propagation based only on colour and spatial proximity can lead to

suboptimal segmentation results. This could be alleviated by the integration of a more sophisticated adaptive semantic segmentation strategy, such as SemanticPaint (Valentin *et al.*, 2015).

Similar to other intrinsic decomposition approaches that rely on user-interaction, this approach assumes that the constraints provided by the user are correct. If the user provides implausible constraints, e.g. paints a constant reflectance stroke across a highly textured region, the optimization will blindly try to satisfy these incorrect constraints thus leading to unrealistic results. Guiding the user and providing some feedback regarding the satisfiability of the provided constraints is a challenging, but interesting, problem for future work.

Finally, this approach is computationally quite demanding and currently requires a state-of-the-art graphics card to achieve real-time performance. A robust, mobile and lightweight solution to the presented problem can be an enabling technology for AR devices.

4.10 Conclusion

A novel real-time approach for user-guided intrinsic decomposition of static scenes is presented. Users can improve the decomposition quality based on live mouse input or an intuitive touch-based interaction metaphor that allows to place decomposition constraints directly in 3D space. The constraints are projected to 2D and used to further constrain the ill-posed intrinsic decomposition problem. The dense reconstruction is also used as a proxy to fuse the obtained reflectance estimates. The novel stabilization term applies constraints based on the projected fused reflectance estimates leading to temporally more coherent decomposition results. The intrinsic decompositions obtained by this approach show state-of-the-art quality at real-time frame rates. In addition, video editing tasks such as recolouring, relighting and material editing are demonstrated based on the obtained decompositions.

While Chapter 3 and Chapter 4 have explored solutions to the intrinsic decomposition problem, they limit the illumination model to a diffuse-surface single-bounce reflection case. While such a simplistic assumption has been shown to produce plausible video editing effects, physically realistic effects require more accurate models of reflectance. In the next chapter (Chapter 5), an extended formulation and technique is presented to solve the intrinsic decomposition problem for not just the first-bounce reflection but also to decompose layers of inter-reflections. Using novel energy terms, this solution leads to more physically accurate decomposition of the scene light transport.

Chapter 5

Live Global Intrinsic Video

Light transport in real world scenes is a complex phenomenon. The light emitted by a light source reflects around the scene many times before it reaches the camera. At each reflection bounce, the light ray loses energy, and also changes its colour. Due to this, a small amount of the multi-bounce light energy received by the camera can have complex colour signature, leading to subtle changes in scene appearance. This chapter proposes a new method that can capture additional lighting layers for multi-bounce reflections from various objects in the scene, in real-time (Meka *et al.*, 2019b). This allows for more physically accurate video editing effects than what was seen in Chapter 3 and Chapter 4. It is shown that such a comprehensive formulation of the intrinsic decomposition problem not only provides additional light transport layers, but also improves the quality of the reflectance and shading layers.

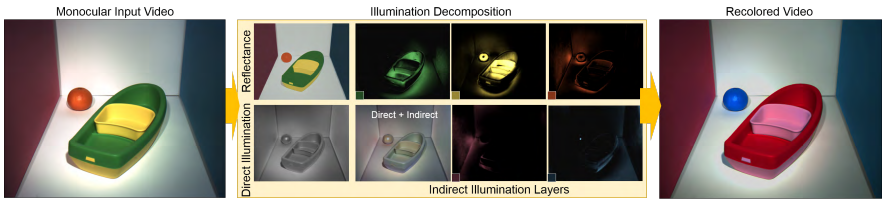


Figure 5.1: The first approach for global intrinsic video decomposition at real-time frame rates is proposed. The approach decomposes a monocular colour video (left) into its reflectance, direct shading, and multiple indirect shading layers (middle) that explain the light transport in the scene up to the first bounce. This enables various live appearance editing applications with interactive user feedback, such as inter-reflection consistent recolouring of objects (right).

5.1 Introduction

Intrinsic image decomposition, as discussed in the earlier chapters, aims to solve the highly challenging, under-constrained problem of decomposing each pixel into the components of light transport, without knowing geometry, light distribution, or materials in a scene. A per-pixel decomposition into intrinsic reflectance (material) and shading layers enables photoreal modification of appearance or illumination in images and videos. However, even the best methods today only manage intrinsic decomposition under starkly simplifying assumptions. Most methods assume Lambertian surfaces and single-bounce local illumination, and therefore model each pixel's colour as a product of one albedo and one shading value. Even under these assumptions, intrinsic decomposition is challenging, and only very few methods achieve temporally coherent, or even real-time video decomposition, as was shown in Chapter 3 and Chapter 4.

Material or shading editing under such two-layer decompositions often fails because global illumination effects make the appearance of various scene points highly inter-dependent, particularly due to phenomena such as inter-reflections, shadows and scattering, see Figure 5.1. Modifying the appearance of one image region may therefore affect many other, possibly more distant, regions in the scene. This inter-dependence of the appearance of surface points cannot be characterized by a single shading layer, but requires a much more challenging decomposition into distinct direct and indirect shading.

Many existing methods that decompose the direct and indirect illumination in a scene depend on active controlled lighting. For instance, sequential illumination of each scene point in Seitz *et al.* (2005) or scene illumination with time-multiplexed patterns in Nayar *et al.* (2006) were used to separate the direct reflection components from the global lighting components. However, these approaches do not capture the appearance inter-dependence between the various points in the scene well, which can be achieved with a different light transport representation (Dong *et al.*, 2015) that enables globally consistent appearance editing. Yet, all these methods are encumbered by complex hardware and acquisition requirements, making it impossible to apply them to normal images or videos.

In contrast, recent image-based methods solve a colour unmixing problem with a sparse set of base colours to decompose an RGB image into layers that can be manipulated independently. Aksoy *et al.* (2016) solve the colour unmixing along with a matting problem without computing interpretable layers such as reflectance or shading. Carroll *et al.* (2011) first compute a two-layer intrinsic image decomposition using the user-interactive method of Bousseau *et al.* (2009), and then solve

the colour unmixing problem on the shading image alone. Both of these methods rely on user input to obtain an initial set of base colours, which is then kept fixed.

Inspired by the sparse base colour assumption, the first real-time method to perform a fully temporally coherent intrinsic decomposition of a video into a reflectance layer, a direct shading layer and multiple indirect shading layers is presented. Decomposition is formulated as a new energy minimization problem that uses layer-specific sparsity priors to estimate per-pixel reflectance, but also direct and indirect shading layers on the basis of concurrently estimated and refined base colours. This proposed formulation employs sparsity of the global layers for effective estimation of the indirect shading layers, for refining the base colours, and for more robust computation of intrinsic reflectance layers. The resulting sparse and dense sets of non-linear equations are separated, and a novel alternating GPU optimization strategy enables efficient computation. The decomposition in the first frame can be optionally refined with a few mouse clicks, which is then automatically propagated to the remaining video. In summary, the core methodical novelties, besides the real-time live system, that distinguish this work from previous literature are:

- Joint decomposition of video frames into a reflectance layer and direct and indirect shading layers, and estimation and refinement of base colours that constitute the scene reflectance.
- A sparsity-based automatic estimation of the underlying reflectance when a user identifies regions of strong inter-reflections.

Based on this decomposition, appearance editing applications are shown on videos, and qualitative and quantitative improvements are demonstrated over the state-of-the-art.

5.2 Related Work

Global Illumination Decomposition To decompose the captured radiance of a scene into direct and indirect components, some methods actively illuminate the scene to investigate the effect of light transport. Seitz *et al.* (2005) use a laser to sequentially light up the corresponding geometry of each pixel, and Nayar *et al.* (2006) and O'Toole *et al.* (2016) use multiple images captured under structured lighting. While these methods use active illumination to decompose scene radiance into direct and indirect components, they cannot separate reflectance and illumination. Thus, these methods cannot ascertain which object causes which colour spill, which makes applications such as recolouring or material editing impossible.

On the other hand, Dong *et al.* (2015) estimate the global illumination caused by diffuse regions of interest, which allows them to perform recolouring on those regions with consistent light interactions with the scene. Laffont *et al.* (2012) present an approach for intrinsic decomposition based on a photo collection of a scene under different viewpoints/illuminations to better constrain the problem. Ren *et al.* (2015) propose a data-driven method for image-based rendering of a scene under novel illumination conditions by taking multiple images of the same scene with different illumination settings as input. Yu *et al.* (1999) estimate the diffuse and specular reflectance map as well as indirect illumination. To this end, they solve inverse radiosity by taking multiple calibrated HDR images with known direct illumination as input along with the geometry of the scene. In contrast, the proposed approach only requires a single colour image or video to estimate the direct reflectance and illumination in addition to decomposing the indirect illumination.

Layer-based Image Editing A physically accurate decomposition is not required to achieve complex image editing tasks such as recolouring of objects. Instead, a decomposition into multiple semitransparent layers is often sufficient, as demonstrated for instance by image vectorization techniques of Favreau *et al.* (2017); Richardt *et al.* (2014). Aksoy *et al.* (2016) introduce an interactive colour unmixing approach that decomposes an image or video into additive layers of dominant scene colours. This enables accurate green-screen keying and layer recolouring, but requires a user to manually identify all base colours. Tan *et al.* (2016) automatically estimate a given number of base colours using the vertices of the simplified convex hull of observed RGB colours. However, the user still needs to determine the order of the layers. Aksoy *et al.* (2017) determine the base colour model fully automatically, and then decompose images into high-quality, additive, near-uniformly coloured layers. They demonstrate a large variety of layer adjustments that are enabled by their decomposition. Innamorati *et al.* (2017) learn an image decomposition into a mixture of additive and multiplicative layers for occlusion, albedo, irradiance and specular layers, instead of layers of distinct colours. Tan *et al.* (2018) perform additive decomposition in real-time given a fixed palette of base colours. In the proposed method, intrinsic decomposition is combined with layer-based decomposition of the shading that enables new video editing tasks that go beyond those supported by existing layer-based decompositions of images.

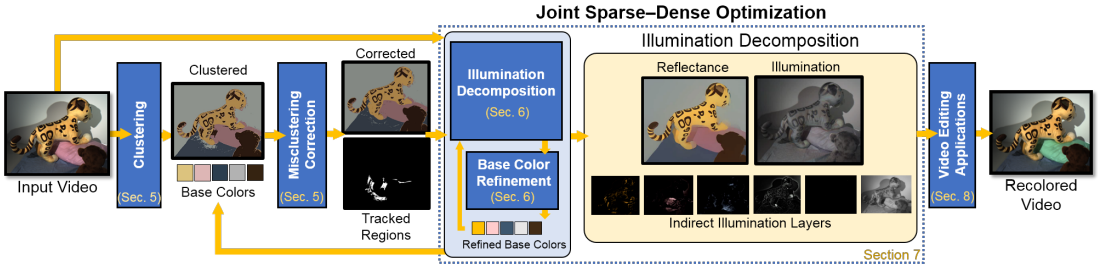


Figure 5.2: Given a monocular colour video as input, the proposed approach estimates the light transport decomposition at real-time frame rates. At the core of this approach are several sparsity priors that enable the estimation of per-pixel direct and indirect shading layers based on a small set of jointly estimated base reflectance colours. The resulting variational problem is efficiently solved using a novel alternating data-parallel optimization strategy. The decomposition is the basis for several compelling live video editing applications, such as inter-reflection consistent recolouring.

5.3 Overview

The first real-time method for temporally coherent intrinsic decomposition of a video into a reflectance layer, direct shading layer and multiple indirect shading layers is presented in this chapter. Figure 5.2 shows the major components of the proposed method and how they interact. A novel sparsity-driven formulation is proposed for the estimation and refinement of a base colour palette, which is used for decomposing the video frames (see Section 5.4). The algorithm starts by automatically estimating a set of base colours that represent scene reflectances (see Section 5.5). Unlike previous methods that heavily rely on user-interaction, the method is automatic and only occasionally requires a minimal set of user clicks on the first video frame to identify regions of strong inter-reflections. The user input is propagated automatically to the rest of the video by a spatio-temporal region-growing method. The intrinsic decomposition and refinement of the base colours (see Section 5.6) is then jointly performed. The proposed formulation results in a mixture of dense and sparse non-convex high-dimensional optimization problems, which are solved efficiently using a custom-tailored parallel iterative non-linear solver that is implemented on the GPU (see Section 5.7). It is shown that this optimization technique achieves real-time frame rates on modern commodity graphics cards.

The proposed method is evaluated on a variety of synthetic and real world scenes. Comparisons show that the method outperforms state-of-the-art intrinsic decomposition and layer-based image editing techniques, both qualitatively and quantitatively (see Section 5.8). It is also demonstrated that real-time global intrinsic decomposition of videos enables a range of advanced, illumination-aware video

editing applications that are suitable for photoreal augmented reality applications, such as inter-reflection-aware recolouring and retexturing (see Section 5.8.4).

5.4 Method

The proposed algorithm decomposes each video frame into a reflectance layer, a direct shading layer and multiple indirect shading layers. First, the algorithm factors each video frame $\mathbf{I}$ into a per-pixel product of the reflectance $\mathbf{R}$ and the shading $\mathbf{S}$:

$$\mathbf{I}(\mathbf{x}) = \mathbf{R}(\mathbf{x}) \odot \mathbf{S}(\mathbf{x}), \tag{5.1}$$

where $\mathbf{x}$ denotes the pixel location and $\odot$ the element-wise product. For diffuse objects, the reflectance layer captures the surface albedo, and the shading layer jointly captures the direct and indirect illumination effects. Unlike most intrinsic decomposition methods, this approach does not use a grayscale shading image, instead the shading layer is represented as a coloured RGB image to allow indirect illumination effects to be expressed in the shading layer.

Inspired by Carroll *et al.* (2011), the shading layer is further decomposed into a grayscale direct shading layer resulting from the white illuminant, and multiple indirect coloured shading layers resulting from inter-reflections from coloured objects in the scene. The first step is estimating a set of base colours that consists of K unique reflectance colours $\{\mathbf{b}_k\}$ that represent the scene. The number K of colours is specified by the user; $K = 10$ is used for all the results, as superfluous clusters will be removed automatically in Section 5.5.1. This set of base colours serves as the basis for the global intrinsic decomposition. The base colours help constrain the values of pixels in the reflectance layer $\mathbf{R}$. For every surface point in the scene, it is assumed that a single indirect bounce of light may occur from every base reflectance colour, in addition to the direct illumination. The global illumination in the scene is modelled using a linear decomposition of the shading layer $\mathbf{S}$ into a direct shading layer T_0 and the sum of the K indirect shading layers $\{T_k\}_{0<k\leq K}$:

$$\mathbf{I}(\mathbf{x}) = \mathbf{R}(\mathbf{x}) \odot \sum_{k=0}^{K} \mathbf{b}_k T_k(\mathbf{x}). \tag{5.2}$$

Here, $\mathbf{b}_0$ represents the colour of the illuminant: white in this case, i.e. $\mathbf{b}_0 = (1,1,1)$. $T_0(\mathbf{x})$ indicates the light transport contribution from the direct illumination. Similarly, the contribution from each base colour $\mathbf{b}_k$ at a given pixel location $\mathbf{x}$ is

measured by the map $T_k(\mathbf{x})$. This scalar contribution, when multiplied with the base colour $\mathbf{b}_k$, provides the net contribution by the base reflectance colour to the global scene illumination. Unlike previous methods, the set of base colours is obtained automatically using a real-time clustering technique. Once the base colours are obtained, the scene clustering can be further refined using a few simple user-clicks. This refines only the regions of clustering but not the base colours themselves.

This specific decomposition assumes that the scene consists of a sparse set of uniformly coloured diffuse objects that are lit by white illumination. It is also assumed that light sources are not visible in the captured videos as they would saturate pixels and hence lead to inaccurate intrinsic decomposition. These simplifying assumptions are also made by the current state-of-the-art approaches such as Carroll *et al.* (2011).

The following sections describe the algorithmic steps to estimate and refine the set of base colours and decompose the input video into the set of global shading layers.

5.5 Base Colour Estimation

The set of base colours is initialized by clustering the dominant colours in the first video frame (Section 5.5.1). This clustering step not only provides an initial base colour estimate, but also a segmentation of the video into regions of approximately uniform reflectance. If needed, the clustering in a video frame undergoes a user-guided correction (Section 5.5.2). The base colours are used for the global intrinsic decomposition (Section 5.6), where they are further refined (Section 5.6.3) and used to compute the direct and indirect shading layers.

5.5.1 Chromaticity Clustering

The first video frame is clustered by colour to approximate the regions of uniform reflectance that are observed in scenes with sparsely coloured objects. The locally constrained clustering approach of Garces *et al.* (2012) segments the image in Lab colour space based on chroma variations using k-means clustering, but has slow, offline run times. In contrast, the proposed approach is based on a much faster histogram-based k-means clustering approach. The clustering of each RGB video frame is performed in a discretized chromaticity space, which makes it more efficient to compute.

The chromaticity image $\mathbf{C}(\mathbf{x}) = \mathbf{I}(\mathbf{x})/|\mathbf{I}(\mathbf{x})|$ is obtained by dividing the input image by its intensity (Bonneel *et al.*, 2014), and then the method previously described in Section 3.4.4 is used to generate a clustered reflectance image. The clustering also produces a segmentation of the input frame, by assigning each pixel

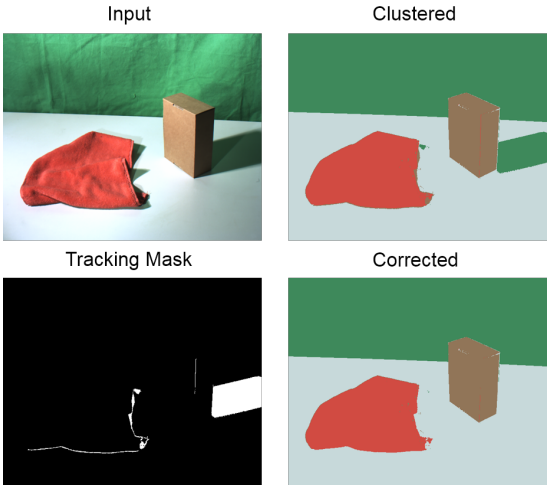


Figure 5.3: Example of misclustering correction (Section 5.5.2). The green colour spill of the background causes misclustered regions in the shadow of the box and the towel (top right). Tracking masks (bottom left) are generated using a few clicks for correcting the misclustered regions (bottom right).

to its closest cluster. This provides a coarse approximation of the reflectance layer, $\mathbf{R}_{\text{cluster}}$, which is used as an initialization for the reflectance layer $\mathbf{R}$ in the energy optimization detailed in Section 5.6.

5.5.2 Misclustering Correction

Since the clustering directly depends on the colour of a pixel, regions of strong inter-reflections may be erroneously assigned to the base colour of an indirect illuminant instead of the base colour representing the reflectance of the region (see the green shadow of the box in Figure 5.3). Such a misclustering is difficult to correct automatically because of the inherent ambiguity of the global intrinsic decomposition problem. In this case, minimal manual interaction is used to identify misclustered regions and then automatically correct the underlying reflectance base colour in all subsequent frames.

5.5.2.1 Region Identification and Tracking

Identifying the true reflectance of a pixel in the presence of strong inter-reflections from other objects is an ambiguous task. In case of direct illumination, the observed colour value of a pixel is obtained by modulating the reflectance solely by the colour of the illuminant (assumed to be white in this case). However, in the case of inter-reflections, there is further modulation by light reflected from other objects, which then depends on their reflectance properties. Such regions are easy to identify by a user, and so the user is asked to simply click on such a region *only in the first frame* it occurs. Then the full region is automatically identified by flood filling using

connected-components analysis based on the cluster identifier. In case the first fill does not cover the full region, additional clicks may be required.

The following method is used for real-time tracking of non-rigidly deforming, non-convex marked regions in subsequent frames. Given the marked pixel region in the previous frame, the same pixel locations are probed in the current frame to identify pixels with the same cluster ID as in the previous frame. Starting from these valid pixels, a flood fill is performed to obtain the tracked marked region in the new frame. To keep this operation efficient, pixels inside the regions are not flood filled. In practice, it is observed that one or two valid pixels are sufficient to correctly identify the entire misclustered region.

5.5.2.2 Reflectance Correction

Once all pixels in a misclustered region are identified in a video frame (either marked or tracked), the sparsity constraint of the indirect shading layers is exploited to solve for the correct reflectance base colour. Multiple full shading decompositions (Section 5.6) are performed for each identified region, evaluating each base colour's suitability as the region's reflectance. For each base colour, the sparsity obtained over the region is measured using the shading sparsity term to be introduced in Equation (5.11). The base colour that provides the sparsest solution of the decomposition is then used as the corrected reflectance. The intuition behind such a sparsity prior is that using the correct underlying reflectance should lead to a shading layer which is explained by the colour spill from only a sparse number of nearby objects, as shown in Figure 5.3.

5.6 Shading Decomposition

Given the initial set of base colours for the scene, the input video is decomposed and the base colours refined. Each input video frame $\mathbf{I}$ is decomposed into its reflectance layer $\mathbf{R}$, its direct shading layer T_0 and a set of indirect shading layers $\{T_k\}$ corresponding to the base colours $\{\mathbf{b}_k\}$ (see Section 5.4). The decomposition into direct and multiple indirect shading layers is inspired by Carroll *et al.* (2011). The direct shading layer T_0 represents the direct contribution to the scene by the external light sources, and the indirect shading layers $\{T_k\}$ capture the inter-reflections that occur within the scene. This decomposition is alternated with the base colour refinement (see Section 5.6.3).

The shading decomposition is formulated as an energy minimization problem with the following energy:

$$E_{\text{decomp}}(\mathbf{X}) = E_{\text{data}}(\mathbf{X}) + E_{\text{reflectance}}(\mathbf{X}) + E_{\text{shading}}(\mathbf{X}), \tag{5.3}$$

where $\mathbf{X} = \{\mathbf{R}, \{T_k\}\}$ is the set of variables to be optimized, while the base colours $\{\mathbf{b}_k\}$ stay fixed. This energy has three main terms: the data fidelity term, reflectance priors (Section 5.6.1) and shading priors (Section 5.6.2). Details on the individual energy terms are given below. This energy is optimized using a novel fast GPU solver (see Section 5.7) to obtain real-time performance.

Data Fidelity Term This constraint enforces that the decomposition result reproduces the input image:

$$E_{\text{data}}(\mathbf{X}) = \lambda_{\text{data}} \cdot \sum_{\mathbf{x}} \left\| \mathbf{I}(\mathbf{x}) - \mathbf{R}(\mathbf{x}) \odot \sum_{k=0}^{K} \mathbf{b}_k T_k(\mathbf{x}) \right\|_2^2, \tag{5.4}$$

where λ_{data} is the weight for this energy term (other terms have their own weights), and the T_k are the $(K+1)$ shading layers of the decomposition: one direct layer T_0, and K indirect layers $\{T_k\}$.

5.6.1 Reflectance Priors

The estimated reflectance layer $\mathbf{R}$ is constrained using three priors:

$$E_{\text{reflectance}}(\mathbf{X}) = E_{\text{clustering}}(\mathbf{X}) + E_{\text{r-sparsity}}(\mathbf{X}) + E_{\text{r-consistency}}(\mathbf{X}). \tag{5.5}$$

The first prior guides the global intrinsic decomposition using the clustered chromaticity map of Section 5.5.1, the second prior encourages a piecewise constant reflectance map using gradient sparsity, and the third prior is a global spatio-temporal consistency prior.

Reflectance Clustering Prior The clustering is used to guide the decomposition, as the chromaticity-clustered image $\mathbf{R}_{\text{cluster}}$ is an approximation of the reflectance layer $\mathbf{R}$. Hence, the reflectance map is constrained to remain close to the clustered image using the following energy term:

$$E_{\text{clustering}}(\mathbf{X}) = \lambda_{\text{clustering}} \cdot \sum_{\mathbf{x}} \|\mathbf{r}(\mathbf{x}) - \mathbf{r}_{\text{cluster}}(\mathbf{x})\|_2^2, \tag{5.6}$$

where the lowercase $\mathbf{r}$ represents the quantity $\mathbf{R}$ in the log-domain, i.e., $\mathbf{r}=\ln\mathbf{R}$, and $\mathbf{r}_{\text{cluster}}$ is the clustered reflectance map.

Reflectance Sparsity Prior Natural scenes generally consist of a small set of objects and materials, hence the reflectance layer is expected to have sparse gradients. Such a spatially sparse solution for the reflectance image can be obtained by minimizing the ℓ_p-norm ($p\in[0,1]$) of the gradient magnitude $\|\nabla\mathbf{r}\|_2$, as was shown previously in Section 3.4.2.

$$E_{\text{r-sparsity}}(\mathbf{X})=\lambda_{\text{r-sparsity}}\cdot\sum_{\mathbf{x}}\sum_{\mathbf{y}\in N(\mathbf{x})}\left\|\mathbf{r}(\mathbf{x})-\mathbf{r}(\mathbf{y})\right\|_2^p, \tag{5.7}$$

where $N(\mathbf{x})$ is the 4-pixel neighbourhood of pixel location $\mathbf{x}$.

Spatiotemporal Reflectance Consistency Prior The spatiotemporal reflectance consistency prior $E_{\text{r-consistency}}(\mathbf{X})$ described in Section 3.4.3 is used here to enforce temporally consistent reflectance. Every pixel is connected with a set of randomly sampled pixels in a small spatiotemporal window and the reflectance of the pixels is constrained to be close under a defined chromaticity-closeness condition.

5.6.2 Shading Priors

The shading $\mathbf{S}$ is constrained to be close to monochrome and the indirect shading layers $\{T_k\}$ to have a sparse decomposition, spatial smoothness and non-negativity:

$$\begin{aligned}E_{\text{shading}}(\mathbf{X})=E_{\text{monochrome}}(\mathbf{X})+E_{\text{i-sparsity}}(\mathbf{X})\\+E_{\text{smoothness}}(\mathbf{X})+E_{\text{non-neg}}(\mathbf{X}).\end{aligned} \tag{5.8}$$

Soft-Retinex Weighted Monochromaticity Prior The shading layer is a combination of direct and indirect illumination effects. Indirect effects such as inter-reflections tend to be spatially local with smooth colour gradients whereas under the white-illumination assumption, the direct bounce does not contribute any colour to the shading layer. Hence, the shading $\mathbf{S}$ is expected to be mostly monochromatic except at small spatial pockets where smooth colour gradients occur due to inter-reflections. Therefore, the following constraint is imposed:

$$E_{\text{monochrome}}(\mathbf{X})=\lambda_{\text{monochrome}}\cdot w_{\text{SR}}\cdot\sum_{\mathbf{x}}\sum_{c}(\mathbf{S}_c(\mathbf{x})-|\mathbf{S}(\mathbf{x})|)^2, \tag{5.9}$$

where $c \in \{R,G,B\}$, and $|\mathbf{S}|$ is its intensity of the shading layer $\mathbf{S}$. This constraint pulls the colour channels of each pixel close to the grayscale intensity of the pixel, hence encouraging monochromaticity. w_{SR} is the soft-colour-Retinex weight computed using

$$w_{\text{SR}} = 1 - \exp(-50 \cdot \Delta\mathbf{C}). \tag{5.10}$$

Here, $\Delta\mathbf{C}$ is the maximum of the chromaticity gradient of the input image in any of the four spatial directions at the pixel location. The soft-colour-Retinex weight is high only for large chromaticity gradients, which represent reflectance edges. Hence, monochromaticity of the shading layer is enforced only close to the reflectance edges and not at locations of slowly varying chromaticity, which represent inter-reflections. Relying on local chromaticity gradients may be problematic when there are regions of uniform coloured reflectance, but in such regions the reflectance sparsity priors tend to be stronger and overrule the monochromaticity prior.

Shading Decomposition Sparsity The shading decomposition is enforced to be sparse in terms of the layers that are activated per-pixel, i.e., those that influence the pixel with their corresponding base colour. Here, the assumption is that during image formation in the real world, a large part of the observed radiance for a scene point comes from a small subpart of the scene. Hence to achieve decomposition sparsity, the sparsity-inducing ℓ_1-norm (Bach *et al.*, 2012) is applied to the indirect shading layers:

$$E_{\text{i-sparsity}}(\mathbf{X}) = \lambda_{\text{i-sparsity}} \cdot \sum_{\mathbf{x}} \left\| \left\{ T_k(\mathbf{x}) \right\}_{k=1}^{K} \right\|_1. \tag{5.11}$$

Spatial Smoothness The decomposition is further encouraged to be spatially piecewise smooth using an ℓ_1-sparsity prior in the gradient domain, similar to Carroll *et al.* (2011), which enforces piecewise constancy of each direct or indirect shading layer:

$$E_{\text{smoothness}}(\mathbf{X}) = \lambda_{\text{smoothness}} \cdot \sum_{\mathbf{x}} \sum_{k=0}^{K} \left\| \nabla T_k(\mathbf{x}) \right\|_1. \tag{5.12}$$

This allows to have sharp edges in the decomposition layers.

Non-Negativity of Light Transport Light transport is an inherently additive process: light bouncing around in the scene adds radiance to scene points, but never subtracts from them. Thus, the quantity of transported light is always positive.

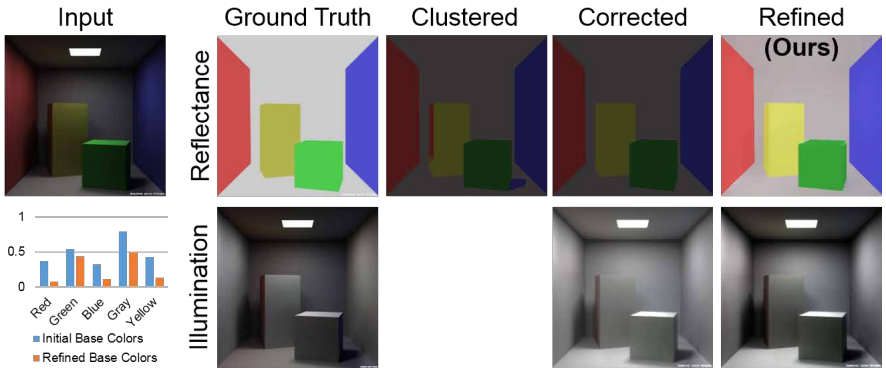


Figure 5.4: Here the improvement obtained by the base colour refinement in the proposed approach is shown. Starting from the clustered reflectance map (Clustered), the misclustering correction step leads to a better segmentation of the scene (Corrected). Finally, the proposed approach fully automatically optimizes for a better set of base colours (Refined). As can be seen, the base colour refinement leads to a significant improvement and results closer to the ground-truth. The bar chart shows the error between the ground-truth base colours and the estimated base colour with (orange) and without (blue) base colour refinement.

Since the shading decomposition layers are motivated by physical light transport, they are enforced to be non-negative to obey this principle:

$$E_{\text{non-neg}}(\mathbf{X}) = \lambda_{\text{non-neg}} \cdot \sum_{\mathbf{x}} \sum_{k=0}^{K} \max(-T_k(\mathbf{x}), 0). \tag{5.13}$$

If the decomposition layer $T_k(\mathbf{x})$ is non-negative, there is no penalty. Otherwise, if $T_k(\mathbf{x})$ becomes negative, a linear penalty is enforced.

5.6.3 Base Colour Refinement

The initial base colours are estimated using chromaticity-based histogram clustering (Section 5.5.1). Unlike previous methods that keep the base colours fixed (Aksoy *et al.*, 2016; Carroll *et al.*, 2011), in the proposed method the base colours are further refined on the first video frame to approach the ground-truth reflectance of the materials in the scene. The refinement of base colours is formulated as an incremental update $\Delta\mathbf{b}_k$ of the base colours $\mathbf{b}_k$ in the original data fidelity term

(Equation (5.4)), along with intensity and chromaticity regularizers:

$$E_{\text{refine}}(\mathbf{X}) = \lambda_{\text{data}} \sum_{\mathbf{x}} \left\| \mathbf{I}(\mathbf{x}) - \mathbf{R}(\mathbf{x}) \odot \sum_{k=0}^{K} (\mathbf{b}_k + \Delta\mathbf{b}_k) T_k(\mathbf{x}) \right\|_2^2 \tag{5.14}$$
$$+ \lambda_{\text{IR}} \sum_{k=1}^{K} \left\| \Delta\mathbf{b}_k \right\|_2^2 + \lambda_{\text{CR}} \sum_{k=1}^{K} \left\| (\mathbf{C}(\mathbf{b}_k) + \Delta\mathbf{b}_k) - \mathbf{C}(\mathbf{b}_k) \right\|_2^2,$$

where $\mathbf{X} = \{\Delta\mathbf{b}_k\}$ is the vector of unknowns to be optimized, λ_{IR} is the weight for the intensity regularizer that ensures small base colour updates, and λ_{CR} is the weight of the chromaticity regularizer, which constrains base colour updates $\Delta\mathbf{b}_k$ to remain close in chromaticity $\mathbf{C}(\cdot)$ to the initially estimated base colour $\mathbf{b}_k$. These regularizers ensure that the base colour update does not lead to oscillations in the optimization process. The refinement energy is solved in combination with the global intrinsic decomposition energy (Equation (5.3)), resulting in an estimation of the unknown variables that together promotes decomposition sparsity. See Figure 5.4 for an example.

This refinement of the base colours leads to a dense Jacobian matrix, because the unknown variables $\{\Delta\mathbf{b}_k\}$ in the energy are influenced by all pixels in the image. This makes the resulting optimization problem difficult to solve in a parallel fashion. The solution to this issue is presented in Section 5.7.

5.6.4 Handling the Sparsity-Inducing Norms

Some energy terms contain sparsity-inducing ℓ_p-norms ($p \in [0, 1]$), i.e., Equations (5.7), (5.11) and (5.12). These objectives are handled in a unified manner using Iteratively Re-weighted Least Squares of Holland & Welsch (1977). To this end, the ℓ_p-norms is approximated by a non-linear least-squares objective based on re-weighting, i.e., the corresponding residuals $\mathbf{r}$ are replaced as follows:

$$\|\mathbf{r}\|_p = \|\mathbf{r}\|_2^2 \cdot \|\mathbf{r}\|_2^{p-2} \tag{5.15}$$
$$\approx \|\mathbf{r}\|_2^2 \cdot \underbrace{\|\mathbf{r}_{\text{old}}\|_2^{p-2}}_{\text{constant}} \tag{5.16}$$

in each step of the applied iterative solver, see also Section 5.7. Here, $\mathbf{r}_{\text{old}}$ is the corresponding residual after the previous iteration step.

5.6.4.1 Handling Non-negativity Constraints

The non-negativity objective in Equation (5.13) contains a maximum function that is non-differentiable at zero. This objective is also handled based on a re-weighting strategy. Thus, the maximum is replaced by a re-weighted least-squares term, $\max(-T_k(\mathbf{x}),0) = w_k T_k^2(\mathbf{x})$, with Here, $\epsilon = 0.002$ is a small constant that prevents division by zero. This transforms the non-convex energy into a non-linear least-squares optimization problem.

5.7 Optimization

The decomposition problems, similar to Chapter 3 and Chapter 4, are all non-convex optimizations based on an objective E with unknowns $\mathbf{X}$. The best decomposition $\mathbf{X}^*$ is obtained by solving the following minimization problem:

$$\mathbf{X}^* = \underset{\mathbf{X}}{\operatorname{argmin}} E(\mathbf{X}). \tag{5.17}$$

The optimization problems are in general non-linear least-squares form and can be tackled by the iterative Gauss–Newton algorithm that approximates the optimum $\mathbf{X}^* \approx \mathbf{X}_k$ by a sequence of solutions $\mathbf{X}_k = \mathbf{X}_{k-1} + \boldsymbol{\delta}_k^*$. The optimal linear update $\boldsymbol{\delta}_k^*$ is given by the solution of the associated normal equations:

$$\boldsymbol{\delta}_k^* = \underset{\boldsymbol{\delta}_k}{\operatorname{argmin}} \left\| \mathbf{F}(\mathbf{X}_{\mathrm{k}-1}) + \boldsymbol{\delta}_k \mathbf{J}(\mathbf{X}_{\mathrm{k}-1}) \right\|_2^2. \tag{5.18}$$

Here, $\mathbf{F}$ is a vector field that stacks all residuals, i.e., $E(\mathbf{X}) = \|\mathbf{F}(\mathbf{X})\|_2^2$, and $\mathbf{J}$ is its Jacobian matrix.

Obtaining real-time performance is challenging even with recent state-of-the-art data-parallel iterative non-linear least-squares solution strategies of Wu *et al.* (2014); Zollhfer *et al.* (2014). To see why this is the case, a closer look at the normal equations is warranted. To avoid cluttered notation, parameters are omitted and simply written as $\mathbf{J}$ instead of $\mathbf{J}(\mathbf{X})$. For the decomposition energies, the Jacobian $\mathbf{J}$ is a large matrix with usually more than 70 million rows and 4 million columns. Previous approaches assume $\mathbf{J}$ to be a sparse matrix, meaning that only a few residuals are influenced by each variable. While this holds for the columns of $\mathbf{J}$ that corresponds to the variables that are associated with the decomposition layers, it does not hold for the columns that store the derivatives with respect to the base colour updates $\{\Delta\mathbf{b}_k\}$, since the base colours influence each residual of E_{data} (Equation (5.4)).

Therefore, $\mathbf{J} = [\mathbf{S_J} \ \mathbf{D_J}]$ has two sub-blocks: $\mathbf{S_J}$ is a large sparse matrix with only a few non-zero entries per row, while $\mathbf{D_J}$ is dense, with the same number of rows, but only a few columns. Thus, the evaluation of the Jacobian $\mathbf{J}$ requires a different specialized parallelization for the dense and sparse parts.

5.7.1 SparseDense Splitting

The described problem is tackled using a sparsedense splitting approach, similar to Section 4.6.2 in the previous chapter, that splits the variables $\mathbf{X}$ into a sparse set $\mathbf{T}$ (decomposition layers) and a dense set $\mathbf{B}$ (base colour updates). Afterwards, $\mathbf{B}$ and $\mathbf{T}$ are optimized for independently in an iterative flip-flop manner. First, $\mathbf{T}$ is optimized, while keeping $\mathbf{B}$ fixed. The resulting optimization problem is a sparse non-linear least-squares problem. Thus, the previous solution is improved upon by performing a non-linear Gauss–Newton step. The corresponding normal equations are solved using 16 steps of data-parallel preconditioned conjugate gradient. The rows of the system matrix are parallelized for using one thread per row (variable).

After updating the 'sparse' variables $\mathbf{T}$, they are kept fixed and the 'dense' variables $\mathbf{B}$ solved. The resulting optimization problem is a dense least-squares problem with a small $3K \times 3K$ system matrix (normally K is between 4 and 7 due to merged clusters). The normal equations are materialized in device memory based on a sequence of outer products, using one thread per entry of $\mathbf{J}^\top\mathbf{J}$. Finally, the system is mapped to the CPU and robustly solved using singular value decomposition. After updating 'dense' variables $\mathbf{B}$, the 'sparse' variables $\mathbf{T}$ are again solved for and this process is iterated until convergence.

5.8 Experiments

The results obtained with the proposed approach are now shown, evaluated qualitatively and quantitatively, and compared to current state-of-the-art decomposition approaches. Please note that the indirect shading layers are scaled for better visualization. The evaluation is performed in terms of robustness, accuracy and runtime on a dataset containing several challenging real and synthetic video sequences. The used test datasets consists of fourteen real and one synthetic sequence (Toys, Box, Boat, Kermit, Cups, Droid, Cart, Girl, Girl2, Umbrella, ChitChat, Hands, Box2 and SyntheticRoom). The proposed approach is compared to the intrinsic decompositions obtained from the approach described earlier in this thesis in Chapter 3 and the state-of-the-art methods of Bonneel *et al.* (2014) and

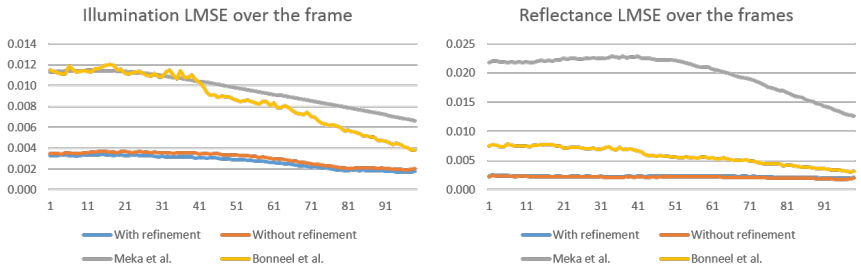


Figure 5.5: The method on the SyntheticRoom sequence is quantitatively analysed. The LMSE error ((Grosse *et al.*, 2009)) is plotted per frame in this graph. The method with base colour refinement achieves the lowest average LMSE score of 0.0024; without, the score is 0.0025, but the result looks visibly worse. two other decomposition techniques are also compared: The real-time intrinsic video method from Chapter 3 has an average error of 0.014, and Bonneel *et al.* (2014) has 0.007.

Carroll *et al.* (2011). The proposed approach is much faster than previous decomposition techniques, and it obtains higher-quality decomposition results in terms of the reflectance map and the indirect shading layers, which directly translates to higher-quality results in all shown applications.

Parameters The following fixed set of parameters are used in all the experiments: $\lambda_{\text{clustering}} = 200$, $\lambda_{\text{r-sparsity}} = 20$, $p = 1$, $\lambda_{\text{i-sparsity}} = 3$, $\lambda_{\text{smoothness}} = 3$, $\lambda_{\text{non-neg}} = 1000$, $\lambda_{\text{data}} = 5000$, $\lambda_{\text{IR}} = 10$, $\lambda_{\text{CR}} = 100$ and $\lambda_{\text{r-consistency}} = \lambda_{\text{monochrome}} = 10$. Since λ_{data} is set to a very high value, the residual of the data term (Equation (5.4)) is below one percent of the intensity range; hence it is too dark to see.

Runtime Performance The performance of this approach is measured on an Intel Core i7 with 2.7 GHz, 32 GB RAM and an NVIDIA GeForce GTX 980. The runtime for videos with a resolution of 640×512 pixels can be broken down into: 14 ms for global intrinsic layer decomposition, 2 s for base colour refinement, and 1 s for misclustering correction. Note that the last two steps, base colour refinement and misclustering correction, are performed only once at the beginning of the video. Afterwards, this approach runs at real-time frame rates ($\geqslant$30 Hz) and enables real-time video editing applications.

5.8.1 Quantitative Results

Quantitative evaluation is performed on the SyntheticRoom sequence. The sequence was rendered using Blender's Cycles renderer. All objects in the scene are assigned diffuse materials, with natural white illumination from the window.

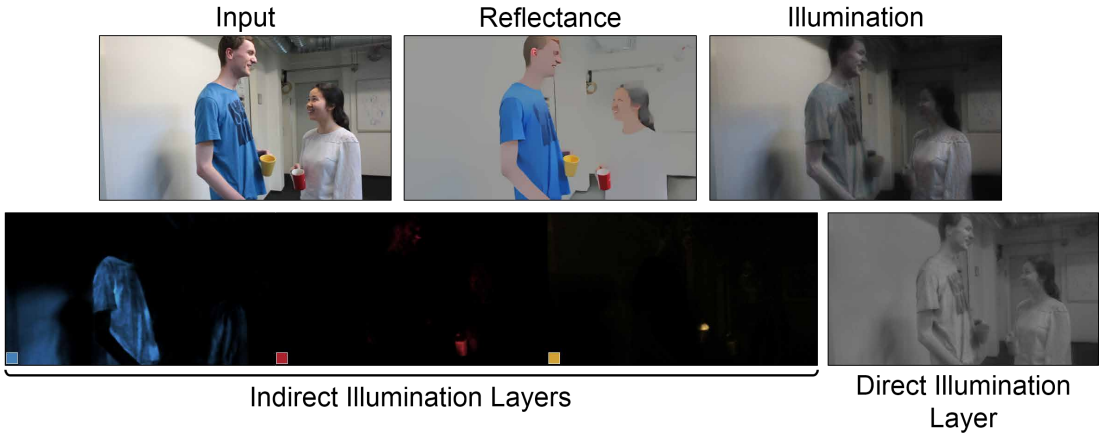


Figure 5.6: Decomposition of the ChitChat sequence. The colour spill is accurately decomposed from the blue shirt and the red cup. Note that the reflectance is devoid of both colour spills.

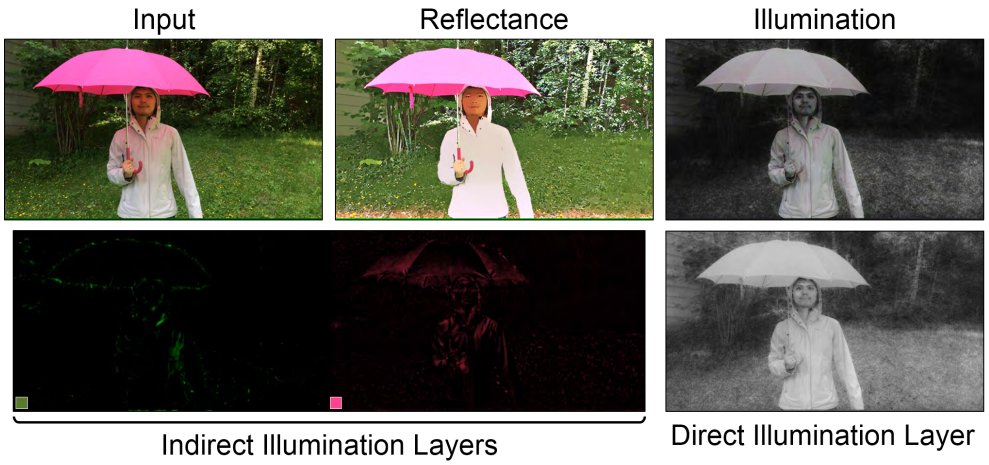


Figure 5.7: Decomposition of the Umbrella sequence. The complex colour spill from the umbrella is mixed with the spill from the forest on the face and the jacket. The method is able to decompose the colours accurately. Note that the reflectance is free from either of the two colour spills, and that both are present in the respective indirect shading layers.

The objects in the scene cause significant inter-reflections. The ground-truth reflectance and shading images are also rendered. The decomposition is compared to the ground-truth using the LMSE error metric proposed by Grosse *et al.* (2009) (Figure 5.5). The error with and without the base colour refinement is analysed, and also compared against state-of-the-art intrinsic video decomposition techniques. The full method obtains the best results.

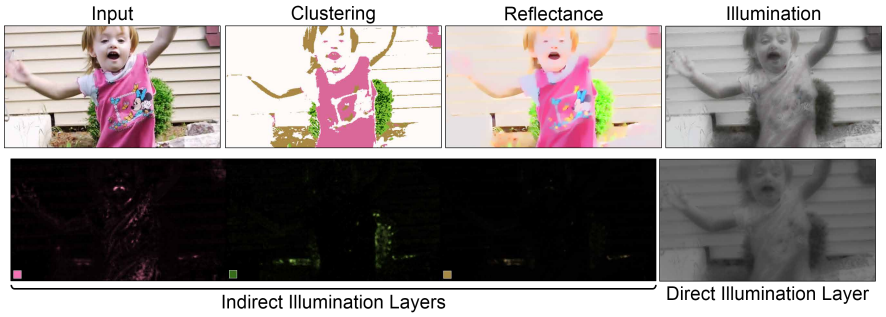


Figure 5.8: The decomposition of the GIRL2 sequence. Even in challenging scenes, where the colour palette is not well defined and thus clustering is difficult, this approach is able to estimate a plausible decomposition along with various indirect shading layers. Note the strong pink inter-reflection on the neck of the girl and within the shirt and in the green bush.

5.8.2 Qualitative Results

It is shown that the indirect shading layers computed by this approach at real-time frame rates nicely capture the inter-reflections between various kinds of objects in a consistent manner, see Figures 5.1 and 5.6 to 5.9. In contrast to intrinsic decomposition approaches, the proposed approach separates the input image into reflectance, coloured direct and indirect shading layers. Please note the colour bleeding of the different parts of the boat in Figure 5.1, which is clearly visible and nicely reconstructed, even though it only accounts for a small amount of the lighting in the input image.

Figure 5.8 shows the global intrinsic decomposition for a complex scene with fast motion and a difficult colour palette. The clustering strategy fails to achieve a meaningful segmentation of the scene. Yet, a plausible decomposition is produced for this challenging scene. In particular, the colour spill from the girl's shirt to her neck and the inter-reflections on the ground from the bush in the background are captured well.

Figure 5.9 shows another example of the reconstructed reflectance and shading layers, where the colour bleeding of the red and blue walls onto the floor is clearly visible. This sequence also shows that the decomposition is temporally coherent and that the shading layers instantly adapt to changes in the scene. Such a decomposition into direct and indirect shading is of paramount importance for illumination-consistent recolouring. An example of this for the same scene is shown in Figure 5.10. Here, the yellow duck is recoloured to purple, which influences the colour of the floor. In another example, the walls are recoloured from blue to red, and vice versa, which also consistently changes the inter-reflections on the floor. Please note that the decomposition is computed at real-time frame rates, which enables the

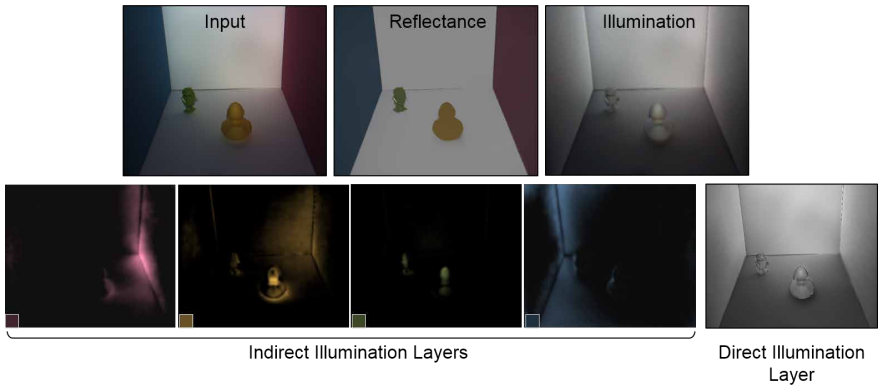


Figure 5.9: Decomposition of the DROID sequence. Note the clean reflectance map and clearly separated colour casts in the indirect shading layers.

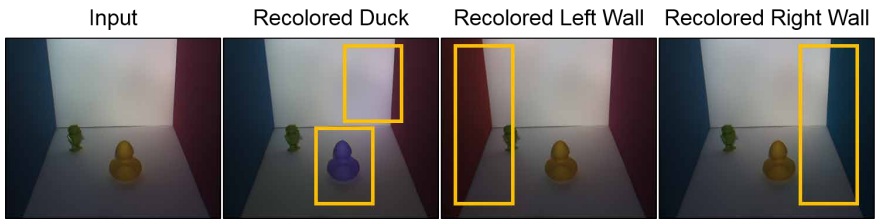


Figure 5.10: The proposed approach enables live recolouring of scene surfaces in a photorealistic and globally consistent manner. Here, the rubber duck and the walls in the scene are recoloured. Note how the corresponding global illumination in the scene (highlighted) is also consistently modified by this approach.

user to explore these effects interactively. In Table 5.1, the number of user-clicks that were performed for each sequence are listed. Please note that most of the sequences did not require user-interaction. Where necessary, only a small number of clicks were required, owing to the region-tracking strategy. To evaluate the method on more general and more complex scenes, which consist of more than just a few prominent objects, the method is tested on images from the *Intrinsic Images in the Wild* dataset of Bell *et al.* (2014). This dataset consists of room-sized indoor scenes. Even though such scenes generally do not exhibit particularly strong global lighting effects, the proposed method is still able to pick up the prominent colours and visualize the global colour spills that occur due to them, as shown in Figure 5.11. Such scenes are challenging for the method to handle, but video editing tasks such as recolouring can still benefit from the decomposition, even in such a challenging setting. A weighted human disagreement rate (WHDR) of 27.2% is obtained, which is better than the baselines and the previously described live intrinsic video method from Chapter 3.

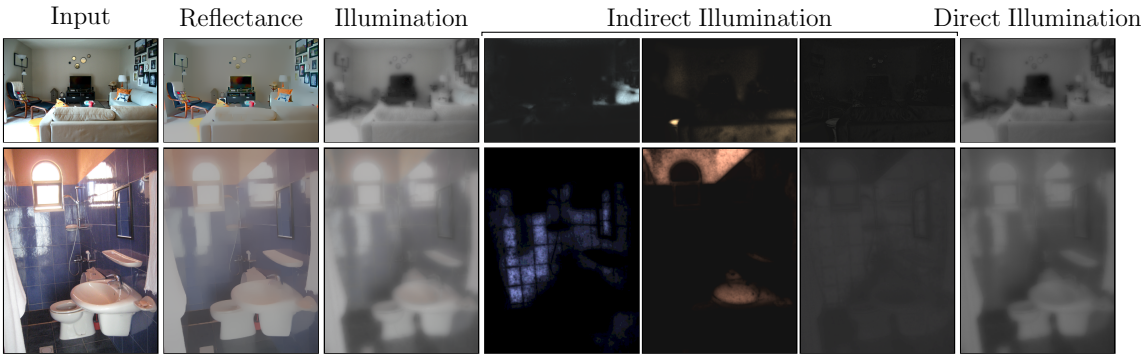


Figure 5.11: Global intrinsic decomposition applied to two samples from the Intrinsic Images in the Wild dataset of Bell *et al.* (2014).

Sequence	Figures	Interactions
Box2		2
Box	Figure 5.3, Figure 5.12	3
Cart	Figure 5.22	1
Cups	Figure 5.20, Figure 5.21	1
Hands		1
Toys	Figure 5.2, Figure 5.13	5
Boat	Figure 5.1	0
ChitChat	Figure 5.6, Figure 5.14	0
Cornell	Figure 5.4, Figure 5.16	0
Droid	Figure 5.9, Figure 5.10, Figure 5.15	0
Girl		0
Girl2	Figure 5.8	0
Kermit	Figure 5.19	0
Paper	Figure 5.17	0
Umbrella	Figure 5.7	0
SyntheticRoom	Figure 5.18	0

Table 5.1: user-interactions required for all sequences. Note that most sequences do not require any user-interaction (bottom half of the table).

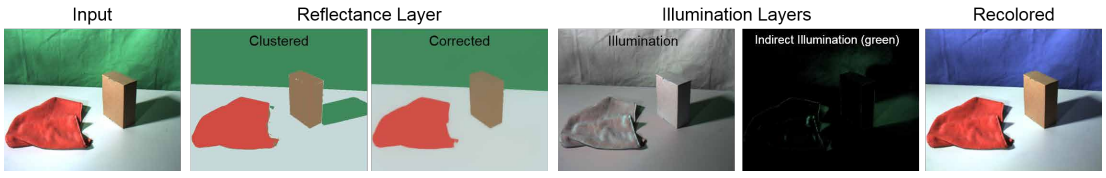


Figure 5.12: Results on the Box sequence, with and without the novel sparsity-based misclustering correction. Regions with strong inter-reflections (shadow of the box) are often misclustered in the reflectance image. This causes indirect illumination to wrongly influence the reflectance layer and not the shading layer, which makes inter-reflection-consistent recolouring impossible. The method alleviates this problem with a little bit of user input to correct the misclustering.

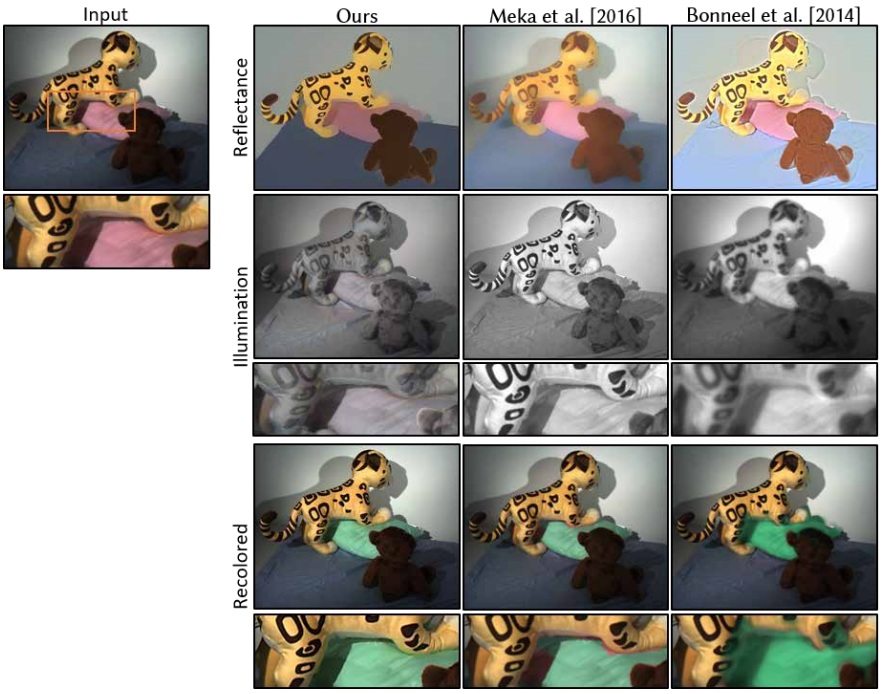


Figure 5.13: Comparison of the proposed global intrinsic decomposition to the live intrinsic video approach described in Chapter 3 (labelled in the figure as Meka *et al.* (2016)) and the state-of-the-art approach of Bonneel *et al.* (2014) on the Toy sequence. With the proposed decomposition, a higher-quality recolouring result than existing methods (see yellow arrows) is achieved. Notice the plausible green colour spill from the pillow onto the toy.

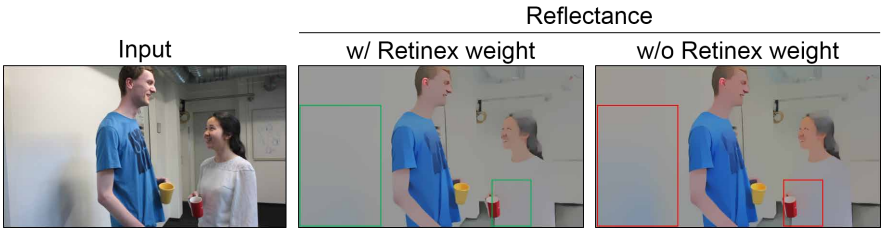


Figure 5.14: Evaluation of the soft-colour-Retinex weight of the monochromatic shading term on the ChitChat sequence. This weight enables the approach to correctly separate the colour spill on the wall and the white shirt.

5.8.2.1 Evaluation of Misclustering Correction

The novel sparsity-based misclustering correction is evaluated in Figures 5.12 and 5.13. In the presence of strong inter-reflections, such as the green colour spill in the shadow of the box in Figure 5.12, estimating the correct reflectance is highly challenging. The state-of-the-art intrinsic decomposition approach described earlier in

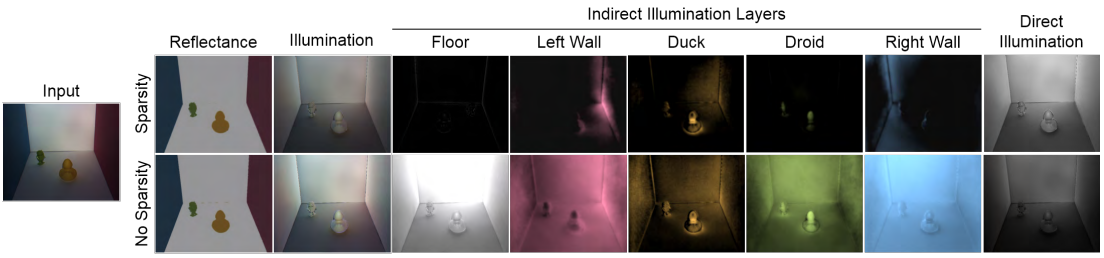


Figure 5.15: Comparison of the global intrinsic decomposition result on the DROID sequence, with and without the shading sparsity prior. Without the sparsity prior, the indirect shading layers, particularly for large regions such as the walls, show activation across the entire image, which is inaccurate. With the sparsity prior, the contribution of the walls to the global illumination is limited to the region close to the walls and in direct sight.

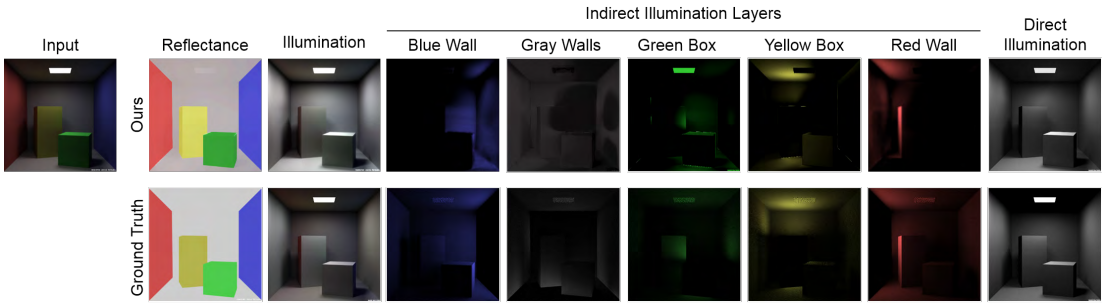


Figure 5.16: The global intrinsic decomposition is qualitatively compared to the ground-truth on the synthetic CORNELL sequence. The estimated indirect shading layers capture the inter-reflections in the scene well. Note that the indirect shading layers are scaled for better visualization.

this thesis (Chapter 3) and that of Bonneel *et al.* (2014) struggle in this scenario, and often miscluster the inter-reflection into the reflectance map, see Figure 5.13. This causes severe problems when an inter-reflection-consistent recolouring of the scene is required, e.g. if the green wall should be virtually replaced by a blue wall. The proposed method alleviates this problem with a minimal amount of user-interaction. With a single click, the misclustered region is identified, and the approach then automatically finds the correct reflectance based on the novel correction strategy that exploits the sparsity of the indirect shading decomposition (see Section 5.5.2). Thus, the reflectance, direct and indirect shading layers computed by the approach enable the seamless inter-reflection-consistent recolouring of scene elements, as shown in Figure 5.12.

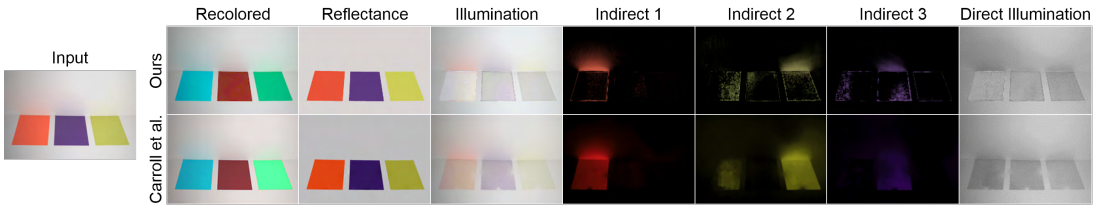


Figure 5.17: Comparison to Carroll *et al.* (2011) on the PAPER sequence. Note that their shading image retains a lot of colour in the coloured paper regions, which results in a direct shading layer that is not uniform across the table and the paper. The base colours ensure that the shading layer retains only the global illumination and not the reflectance. This results in sparser shading layers, while accurately representing the colour spill from the paper.

5.8.2.2 Evaluation of the Sparsity Prior

The importance of the sparsity prior is evaluated in Figure 5.15 by comparing the global intrinsic decomposition result with and without the shading sparsity prior (Equation (5.11)). Without the sparsity prior, the indirect shading layers show activations across the entire image domain, which is inaccurate. The sparsity prior forces inter-reflections to be explained by a small number of base colours; thus the optimization has to choose how to optimally explain the inter-reflections. This leads to sparser and more realistic indirect shading layers that enable accurate inter-reflection-consistent recolouring. Note that with the sparsity prior as expected from physical light transport the contribution of the walls to the global illumination is limited to the regions close to the walls and in direct sight.

5.8.2.3 Evaluation of the Soft-Colour-Retinex Weight

The importance of the soft-colour-Retinex weight in the shading monochromaticity prior is evaluated in Figure 5.14. Without the soft-Retinex weight, the prominent blue colour spill on the wall and the red spill on the white shirt are both incorrectly contained in the reflectance layer. This problem is easily resolved by the soft-Retinex weight.

5.8.3 Comparisons

A ground-truth comparison on synthetic data is shown in Figure 5.16. In the following, the proposed approach is compared to the state-of-the-art decomposition technique described earlier in Chapter 3 and the approaches of Carroll *et al.* (2011) and Bonneel *et al.* (2014).

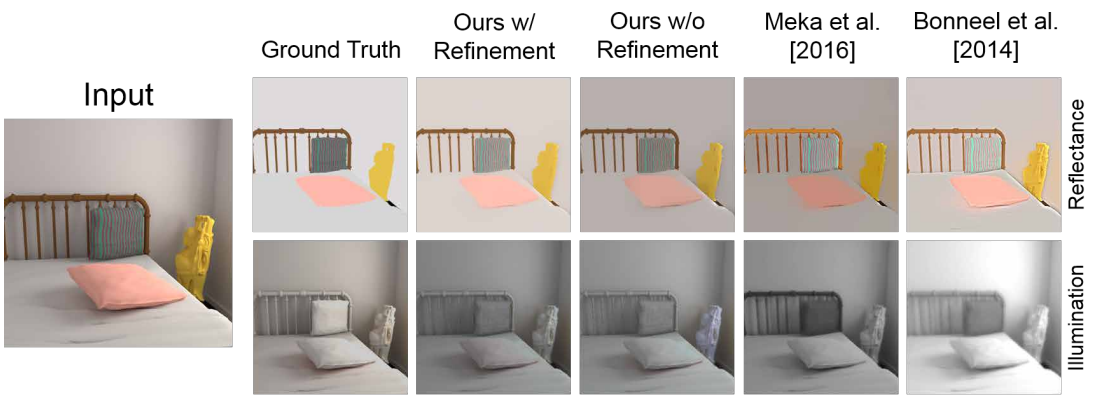


Figure 5.18: Decomposition result on the SYNTHETICROOM sequence. Without the base colour refinement, the proposed approach incorrectly estimates blue shading on the statue, the bedpost and the cushion. The result with the base colour refinement is closer to the ground-truth. The colour-spill from the pink cushion and within the yellow statue is accurately decomposed into the shading layer, while the intrinsic video decomposition methods incorrectly bake them into the reflectance layer.

5.8.3.1 Comparison to Carroll *et al.* (2011)

Their results in Figure 5.17 retain too much colour in the coloured paper regions of the indirect shading layers (Figure 5.17, bottom), resulting in a direct shading layer that is not uniform across the table and the papers. The base colour refinement ensures that the shading image retains only the global illumination (Figure 5.17, top), and that the colour variation that stems from actual surface reflectance variation is moved to the reflectance layer. This causes the shading layers to be more sparse, while accurately representing the colour spill from the paper. Note that these results are obtained automatically, while Carroll *et al.*'s approach requires several user scribbles.

5.8.3.2 Comparison to Bonneel *et al.* (2014) and Meka *et al.* (2016) (the approach from Chapter 3)

In Figure 5.18, the base colour refinement strategy is analysed on a synthetic sequence. Without the refinement, the shading is inaccurately estimated to be blueish in multiple places, which is resolved by the refinement strategy. The other methods obtain globally inconsistent shading results, and incorrectly bake the colour spills into the reflectance layer. In Figure 5.19, the proposed approach is compared to the live intrinsic video decomposition approach from Chapter 3. Their approach does not correctly handle inter-reflections, while the proposed approach enables inter-reflection-consistent recolouring of scene objects. Please note the

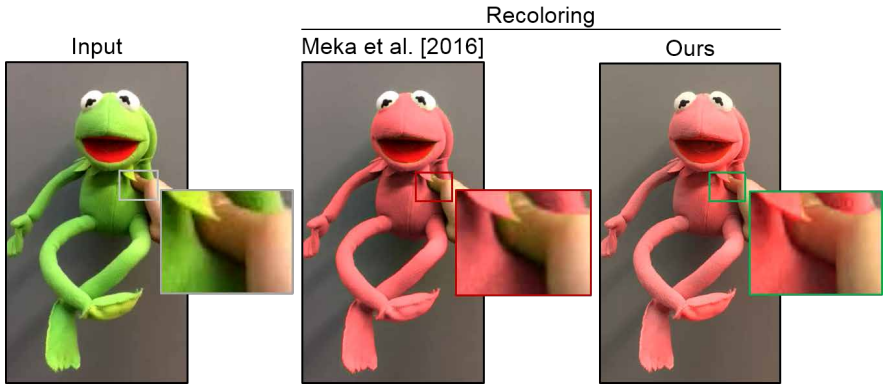


Figure 5.19: Comparison of recolouring results to the previously described approach from Chapter 3 (labelled in the figure Meka *et al.* (2016)) on the KERMIT sequence from Bonneel *et al.* (2014). The earlier approach approach does not correctly handle inter-reflections, e.g. from Kermit onto the thumb, while the approach consistently reconstructs and recolours these inter-reflections.

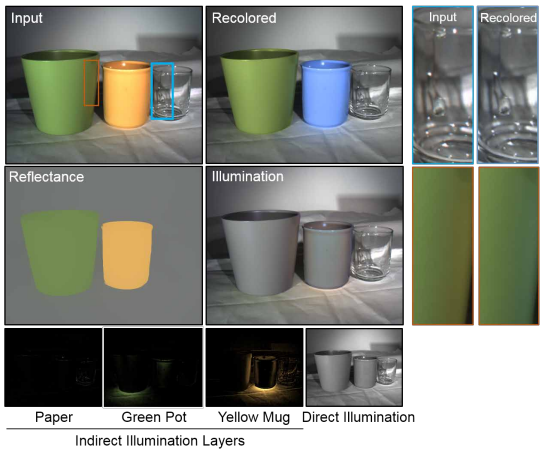


Figure 5.20: Recolouring result on the CUP sequence]Recolouring result on the CUP sequence. Apart from the prominent colour spills on the table cloth, even subtle inter-reflections on the green pot and the glass are captured well.

colour bleeding from the green frog onto the hand. A second comparison is shown in Figure 5.13, where the proposed approach is compared to the offline intrinsic video decomposition approach of Bonneel *et al.* (2014). Neither of these methods is able to correctly handle scene inter-reflections.

5.8.4 Interactive Live Applications

Several live video applications are demonstrated based on the proposed global intrinsic decomposition approach, such as inter-reflection-consistent recolouring and colour keying. For a survey of digital keying methods, refer to Schultz & Hermes (2006).

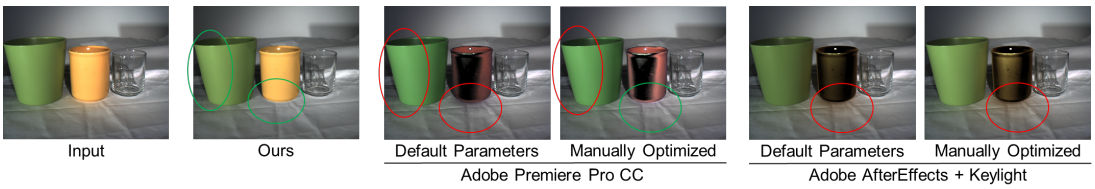


Figure 5.21: Comparison to recolouring software. The commercial software cannot remove the colour spill for a particular object, such as the yellow cup. It only supports removing a particular colour component completely from the entire image. Results are shown with the default parameters and manually tuned parameters for the best results. Even with manually tuned parameters, the software packages cannot coherently deal with the colour spill and lead to artifacts or inaccuracies.

5.8.4.1 Inter-Reflection-Consistent Recolouring

The proposed global intrinsic decomposition approach enables inter-reflection-consistent recolouring of live video streams. An object can be recoloured by modifying its associated base colour, which consistently recolours the objects reflectance and indirect shading layer. Several plausible inter-reflection-consistent recolouring results are already shown in Figures 5.10, 5.12, 5.13 and 5.19, which outperform existing intrinsic image decomposition approaches of (Bonneel *et al.*, 2014; Meka *et al.*, 2016). In Figure 5.20, it is further demonstrated that the proposed approach can even recolour subtle inter-reflections on glass, and not just on diffuse surfaces.

5.8.4.2 Inter-Reflection-Consistent Colour Keying

Colour keying is a technique often used in visual effects for overlaying a subject in a video on top of a different background using a colour-based segmentation. In practice, a uniform green background is often used. Global light transport in the scene often causes green inter-reflections from the background onto the subject. This leads to unrealistic composites, since a green colour spill is often visible on the subject, which does not match the new background. The proposed interactive global intrinsic decomposition approach can be used to alleviate this problem, as shown in Figure 5.12. The input video is first separated into its reflectance, direct and indirect shading components. Afterwards, the base colour of the green indirect shading layer is modified, which relights the subject to better match the new background. This leads to more realistic outputs and can be achieved at interactive frame rates with the proposed approach.

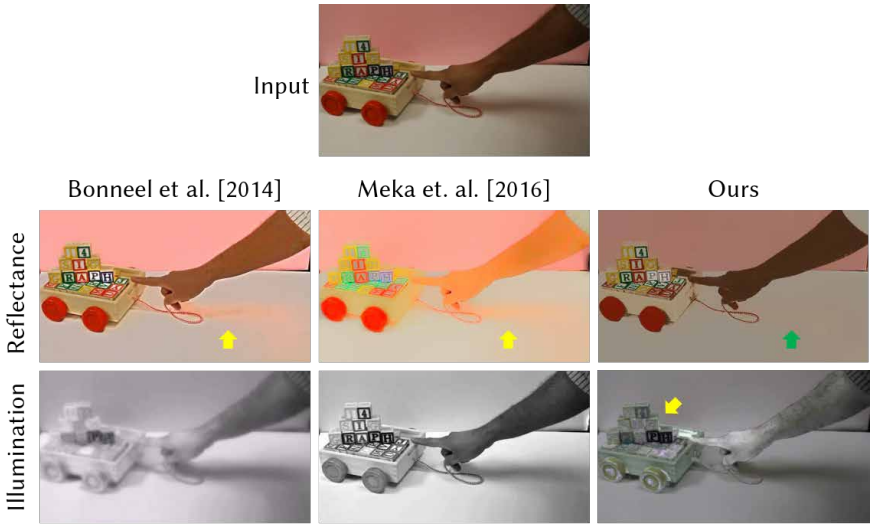


Figure 5.22: Comparison to state-of-the-art techniques on the CART sequence. Note the shadow of the hand and the resulting inter-reflections on the table. This technique correctly places the inter-reflections into the shading layer, while they are baked into the reflectance layer for the other methods, including the one from Chapter 3 (labelled in the figure Meka *et al.* (2016)). However, due to the large number of base colours in the scene, the proposed method incorrectly decomposes the reflectance and shading for the blocks.

5.8.4.3 Colour-Spill Suppression

In many video editing tasks, suppressing a strong colour spill is highly important. This technique is often used in movie and television productions to suppress the spill from a green or blue-screen. An example of such an application is shown in Figure 5.21. The spill from the shiny yellow cup is successfully suppressed by removing the indirect shading layer of the cup from the global intrinsic decomposition and recombining the other layers. The results are compared with state-of-the-art commercial software. The tested software is not able to suppress the spill for a particular object, but only for a particular colour scheme. The parameters of the software were also manually tuned to achieve the best results. After optimizing the parameters, Adobe Premiere Pro CC is able to suppress the spill from the cup, but it also incorrectly modifies the colour of the green cup on the left side. As is evident, the proposed approach achieves the best results.

5.9 Discussion

While high-quality global intrinsic decomposition results and a wide range of applications have been demonstrated, the proposed approach still has some limitations that may be addressed in follow-up work. The proposed approach only has limited scene information available and thus cannot model parts of the scene that are outside the view of the camera. This means that inter-reflections caused by out-of-view objects cannot be properly modelled, since the corresponding base colour might not be available. This is a common limitation of all illumination decomposition approaches, including Carroll *et al.* (2011).

A further restriction is the user-specified upper bound on the number of base colours. If an object with an unseen colour enters the scene for the first time, and the base colours are already exceeded, its inter-reflections cannot be modelled. This limitation could be alleviated in the future with a dynamic clustering strategy. Quick changes in camera view or abrupt scene motion can break the region propagation strategy. This could be alleviated by more sophisticated tracking strategies, such as SLAM.

Complex, textured scenes with many different colours are challenging to decompose, e.g. see Figure 5.22, since this requires many base colours, leading to a large number of variables and an even more under-constrained optimization problem. More sophisticated potentially learned scene priors could be beneficial. The proposed approach only obtains plausible decompositions for first bounce reflections. Modelling the higher-order bounces would require a dramatic increase in the number of base colours, since all mixtures of reflectances would have to be considered. More general indoor and outdoor scenes such as those in the *Intrinsic Images in the Wild* dataset of Bell *et al.* (2014) are not the ideal use cases for this method. This is because the scene illumination is often extremely complex, e.g., due to coloured light sources and tinted windows. Like most approaches, this one assumes white direct illumination. Dealing with coloured light sources is a more challenging problem due to the larger number of variables and thus greater ambiguity in the decomposition. Yet, assuming some level of sparsity in the colour of the light sources, the problem could still be solved using a similar formulation. This would be a very interesting direction for future work.

5.10 Conclusion

This chapter presents the first global intrinsic decomposition approach for videos. At the core of the proposed method are multiple interlinked energies that enable the estimation of the direct and indirect decomposition layers based on a small set of jointly estimated base colours. Decomposition results that qualitatively improve on existing state-of-the-art methods in addition to various compelling appearance editing applications have been demonstrated.

This part of the thesis has presented novel solutions to the real-time intrinsic scene decomposition problem, but was limited to work with diffuse surfaces. The next part of the thesis focuses on relaxing this assumption and working with a more general class of reflectance surfaces that exhibit more complex effects such as specular reflections and sub-surface scattering, particularly for the challenging case of acquiring surface reflectance properties in *real-time*.

Part II

Real-time Reflectance Acquisition

Chapter 6

Live Intrinsic Material Estimation

The majority of the materials encountered in day-to-day life exhibit complex reflectance effects. While the real-time intrinsic decomposition methods that were developed in Part I work well at scene level, modelling and acquiring reflectance at object level requires a more sophisticated formulation. Part II of the thesis deals with acquiring the reflectance of non-diffuse surfaces in real-time. In particular, this chapter presents the first method to estimate in real-time the reflectance of a uniform material object of arbitrary shape from a monocular RGB image (Meka *et al.*, 2018). This allows for the first time to demonstrate new augmented reality applications such as material cloning and transfer.

6.1 Introduction

The estimation of material properties from a single monocular colour image is a high-dimensional and under-constrained problem. The blind deconvolution nature of the problem has attracted usage of complex setups and, more recently, various natural and handcrafted priors, but has yet remained outside the scope of real-time implementation due to the resulting dense optimization problem. Given these challenges, previous real-time approaches such as the ones described earlier in this thesis in Part I and other state-of-the-art methods like Mandl *et al.* (2017), have predominantly focused on estimating diffuse materials. In this work, a much harder inverse problem is tackled by additionally estimating specular material properties, such as specular colour and material shininess, as well as segmentation masks for general objects of uniform material from a single colour image or video in real-time.

Recent advances in deep learning enable the automatic learning of underlying natural subspace constraints directly from large training data, while also reducing

the need to solve the expensive dense non-linear optimization problem directly. Some recent work has successfully demonstrated the capability of convolutional neural networks to solve the inverse rendering problem of separating material from illumination, particularly in the context of uniform material objects. Current approaches estimate material from one (Georgoulis *et al.*, 2017b; Liu *et al.*, 2017) or more images (Kim *et al.*, 2017; Wu *et al.*, 2016). Georgoulis *et al.* (2017b) learn BRDF parameters and outdoor environment maps from single images of specular objects from a specific class (cars, chairs or couches only). Kim *et al.* (2017) estimate BRDF parameters from multiple RGB input images in 90 ms. Shi *et al.* (2017) perform intrinsic image decomposition of a single object image into diffuse and specular layers but do not solve the denser and more complex material estimation problem.

Most of these methods take a direct approach to parameter regression without any additional supervision, due to which the network may not necessarily learn to perform the physical deconvolution operation that is intrinsic to inverse rendering, and hence runs the risk of simply overfitting to the training data. The exception is the approach of Liu *et al.* (2017) that took a first important step in this direction using an expert-designed rendering layer. However, such a rendering layer requires shape estimation in the form of surface normals, which are challenging to regress for general objects. This limits their method to objects of particular shape classes (cars, chairs and couches), and also requires manual segmentation of the object in the image.

In contrast, the proposed approach is the first real-time material estimation method that works for objects of any general shape, and without manual segmentation, making it applicable to live application scenarios. The proposed approach draws inspiration from the comptuer graphics rendering process. The input image is decomposed into intrinsic image layers and fine-grained intermediate supervision is provided by following the rendering process closely. The task of material estimation is decoupled from the shape estimation problem by introducing a novel image-space supervision strategy on the intrinsic layers using a highly efficient perceptual loss that makes direct use of the regressed layers. Finally, each material parameter is regressed from the relevant intrinsic image layers, self-supervised by the perceptual rendering loss. This mechanism results in a demonstrably more accurate material estimation.

In addition to these core innovations, this approach distinguishes itself from previous work in the following ways:

- It fully automatically performs object segmentation in the image, enabling the method to be applied to single images and live videos.

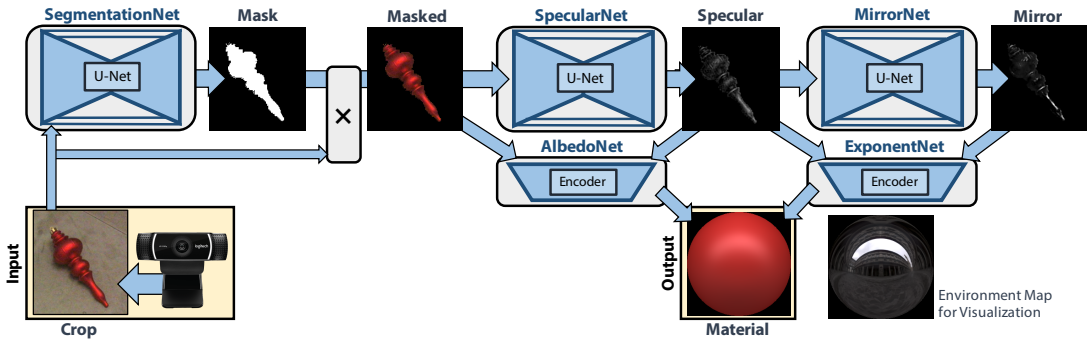


Figure 6.1: The proposed approach enables real-time estimation of material parameters from a single monocular colour image (bottom left). The proposed end-to-end learning approach decomposes the complex inverse rendering problem into sub-parts that are inspired by the physical real world image formation process, leading to five specifically tailored subnetworks. The complete network is trained in an end-to-end fashion. Environment map used for material visualization (bottom right) courtesy of Debevec (1998).

- The networks are trained for the challenging indoor setting, and successfully handle complex high-frequency lighting as opposed to the natural outdoor illumination used by other methods such as Georgoulis *et al.* (2017b); Kim *et al.* (2017); Liu *et al.* (2017), since most mixed-reality applications are used indoors.
- If shape information is available, e.g. from a depth sensor, the proposed method also extracts separate low- and high-frequency lighting information, which is crucial for vivid AR applications.

6.2 Related Work

The appearance of an object in an image depends on its surface geometry, material and illumination. Estimation of these components is a fundamental problem in computer vision, and joint estimation the ultimate quest of inverse rendering (Ramamoorthi & Hanrahan, 2001c; Yu *et al.*, 1999). Geometry reconstruction has seen major advances since the release of commodity depth sensors (e.g. Dai *et al.*, 2017; Innmann *et al.*, 2016; Izadi *et al.*, 2011; Niener *et al.*, 2013; Zollhfer *et al.*, 2015). However, estimation of material and illumination remains relatively more challenging. Approaches for estimating material and illumination need to make strong assumptions, such as the availability of a depth sensor (Guo *et al.*, 2017; Rohmer *et al.*, 2017), lighting conditions corresponding to photometric stereo (Hui & Sankaranarayanan, 2017), a rotating object under static illumination (Xia *et al.*, 2016), multiple images

of the same object under varying illumination (Wang *et al.*, 2018), having an object of a given class (Georgoulis *et al.*, 2017b), or requiring user input (Meka *et al.*, 2017b).

Material Estimation There are broadly two classes of material estimation approaches: (1) approaches that assume known geometry, and (2) approaches for specific object classes of unknown geometry. Methods that require the surface geometry of objects to be known can, in principle, work on any type of surface geometry. Dong *et al.* (2014) estimate spatially-varying reflectance from the video of a rotating object of known geometry. Wu & Zhou (2015) perform on-the-fly appearance estimation by exploiting the infrared emitterreceiver system of a Kinect as an active reflectometer. Knecht *et al.* (2012) also propose a method for material estimation at interactive frame rates using a Kinect sensor. Li *et al.* (2017) learn surface appearance of planar surfaces from single images using self-augmented CNNs. There are also several recent offline methods such as Kim *et al.* (2017); Richter-Trummer *et al.* (2016); Wu *et al.* (2016) that capture a set of RGB images along with aligned depth maps to estimate an appearance model for the surface geometry. Recent methods by Rematas *et al.* (2016), Georgoulis *et al.* (2017b) and Liu *et al.* (2017) do not assume known geometry, but instead rely on implicit priors about the object shape, and therefore only work on the specific classes of objects such as cars or chairs for which the methods are trained.

In contrast to these methods, the proposed approach requires neither known surface geometry nor is it restricted to specific object classes. To the best of my knowledge, the only other RGB-only method that works on arbitrary objects is by Lombardi & Nishino (2016b). However, it is an offline method. The proposed real-time method can significantly enhance a wide variety of applications like material editing similar to Carroll *et al.* (2011); Di Renzo *et al.* (2014); Dong *et al.* (2015); Khan *et al.* (2006), object relighting as in Liao *et al.* (2015), cloning and insertion.

Illumination Estimation Assuming a diffuse reflectance, Marschner & Greenberg (1997) estimate environment maps from captured RGB images and scanned geometry. Given a single input image, methods exist for estimating natural outdoor illumination (Hold-Geoffroy *et al.*, 2017; Lalonde *et al.*, 2012), indoor illumination (Gardner *et al.*, 2017) or the location of multiple light sources (Lopez-Moreno *et al.*, 2013). Georgoulis *et al.* (2017a) estimate an environment map from the photo of a multicoloured specular object of known shape. Mandl *et al.* (2017) similarly learn the lighting from a single image of a known object. Lalonde & Matthews (2014) perform illumination estimation from an image collection used for structure-from-motion reconstruction. However, note that the main contribution of the proposed

method lies in material estimation, and not illumination estimation. Nevertheless, given geometry, it is shown in Section 6.8 how this approach can be extended to additionally estimate illumination.

6.3 Overview

The proposed approach is the first end-to-end algorithm for real-time estimation of an object's material and segmentation mask from just a single colour image. The image formation model is presented in Section 6.4. In Section 6.5, the tackling of the underlying inverse rendering problem using encoderdecoder architectures (Isola *et al.*, 2017; Ronneberger *et al.*, 2015) is discussed. For higher temporal stability, when the method is applied to video, the reconstructed material parameters are temporally fused. This is discussed in Section 6.6. The results are evaluated and compared to state-of-the-art techniques in Section 6.7. Finally, in Section 6.8, mixed-reality applications are demonstrated that benefit from the proposed real-time inverse rendering approach, such as seamless placement of virtual objects, with real-time captured materials, in real world scenes.

6.4 Appearance Model

In this section, the forward process of image formation and all employed scene assumptions are explained. The appearance of an object in an image depends on its bidirectional reflectance distribution function (BRDF) and the light transport in the scene. The light transport is modelled based on the trichromatic approximation of the rendering equation (Equation (2.1)) described earlier in Section 2.1.

To make real-time inverse rendering tractable, a few simplifying assumptions are made, which are widely used, even in the offline state-of-the-art inverse rendering techniques of Georgoulis *et al.* (2017b); Liu *et al.* (2017). First, it is assumed that the object is not emitting light, i.e., it is not a light source, and only direct illumination is modelled. Global changes in scene brightness are modelled based on an ambient illumination term $\mathbf{L}_a \in \mathbb{R}^3$. Distant lighting and the absence of self-shadowing is assumed, which decouples the incident illumination from the object's spatial embedding. Given these assumptions, the rendering equation simplifies to

$$\mathbf{L}(\mathbf{x},\boldsymbol{\omega}_o) = \mathbf{L}_a + \int_{\Omega} \underbrace{\mathbf{f}(\mathbf{x},\boldsymbol{\omega}_i,\boldsymbol{\omega}_o)(\boldsymbol{\omega}_i \cdot \mathbf{n})}_{\mathbf{BP}(\mathbf{x},\mathbf{n},\boldsymbol{\omega}_i,\boldsymbol{\omega}_o)} \mathbf{E}(\boldsymbol{\omega}_i) d\boldsymbol{\omega}_i. \tag{6.1}$$

Distant illumination is represented using an environment map $\mathbf{E}(\boldsymbol{\omega}_i)$. Diffuse and specular object appearance is parameterized using the BlinnPhong reflection model of Blinn (1977) described earlier in Section 2.2:

$$\mathbf{BP}(\mathbf{x},\mathbf{n},\boldsymbol{\omega}_i,\boldsymbol{\omega}_o) = \underbrace{\mathbf{m}_d(\boldsymbol{\omega}_i \cdot \mathbf{n})}_{\text{diffuse}} + \underbrace{\mathbf{m}_s(\mathbf{h} \cdot \mathbf{n})^s}_{\text{specular}}. \tag{6.2}$$

Here, $\mathbf{m}_d \in \mathbb{R}^3$ is the diffuse, and $\mathbf{m}_s \in \mathbb{R}^3$ the specular material colour (albedo). Note that a white specularity is assumed, i.e., $\mathbf{m}_s = \alpha \cdot \mathbf{1}_3$, with $\mathbf{1}_3$ being a 3-vector of ones. The halfway vector $\mathbf{h} = \frac{\boldsymbol{\omega}_i + \boldsymbol{\omega}_o}{\|\boldsymbol{\omega}_i + \boldsymbol{\omega}_o\|}$ depends on the light direction $\boldsymbol{\omega}_i$ and the viewing direction $\boldsymbol{\omega}_o$. The scalar exponent $s \in \mathbb{R}$ determines the size of the specular lobe, and thus the shininess of the material.

6.5 Learning

The goal of the proposed approach is the real-time estimation of diffuse and specular object material from a single colour image. This high-dimensional and non-linear inverse rendering problem is ill-posed, since each single colour measurement is the integral over the hemisphere of the product between the BRDF and the incident illumination modulated by the unknown scene geometry (see Equation (2.1)).

A novel discriminative approach is proposed to tackle this challenging problem using deep convolutional encoderdecoder architectures. In the following, the synthetic ground-truth training corpus, the physically motivated inverse rendering network and a novel perceptual per-pixel rendering loss are described, and it is shown how the entire network can be trained end-to-end.

6.5.1 Synthetic ground-truth Training Corpus

Since the annotation of real world images with ground-truth BRDF parameters is practically infeasible, the deep networks are trained on fully synthetically generated imagery with readily available ground-truth. The training corpus

$$\mathbf{T} = \{\mathbf{I}_i, \mathbf{B}_i, \mathbf{D}_i, \mathbf{S}_i, \mathbf{M}_i, \mathbf{BP}_i\}_{i=1}^{N}$$

consists of $N = 100{,}000$ realistically rendered images $\mathbf{I}_i$, their corresponding binary segmentation masks $\mathbf{B}_i$, diffuse shading images $\mathbf{D}_i$, specular shading images $\mathbf{S}_i$, mirror images $\mathbf{M}_i$, and the ground-truth BlinnPhong parameters $\mathbf{BP}_i$. See Figure 6.2 for examples from the training corpus.

Figure 6.2: Two examples from the synthetic ground-truth training corpus (from left to right): colour image **I**, segmentation mask **B**, diffuse **D** and specular **S** shading image, mirror image **M**, and environment map **E**.

Each of the N training frames shows a single randomly sampled object from a set of 55 synthetic 3D models[1,2] (50 models for training and 5 for testing). The object is rendered with random pose, orientation, size and BlinnPhong parameters $\mathbf{BP}_i$ using perspective projection to obtain the training corpus **T**. The albedo parameters are sampled uniformly in the YUV colour space and then converted to RGB.

The object is lit with a spherical environment map $\mathbf{E}_i$, which is randomly sampled from a set of 45 indoor maps that were captured with an LG 360 Cam with manual exposure control, see Figure 6.2 (right). The environment maps were captured in varied indoor settings, in rooms of different sizes and different lighting arrangement, such as homes, offices, classrooms and auditoriums. For data augmentation, the environment maps are randomly rotated with the condition that there is a strong light source in the frontal hemisphere. This ensures that highlights will be visible if the object is specular.

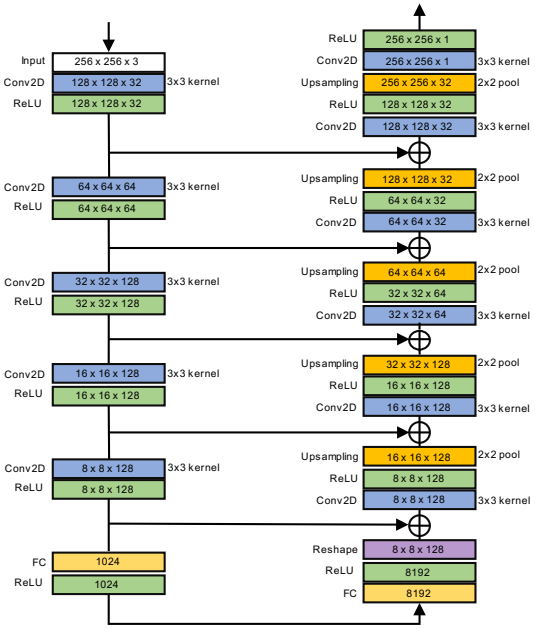
Objects are rendered under different perspective views and crops are obtrained around the objects at different resolutions with varying amounts of translation and scaling. A background is added based on random textures to the rendered object image to provide sufficient variety for the segmentation network to learn foreground segmentation. The training corpus is made publicly available.[3]

6.5.2 Physically Motivated Network Architecture

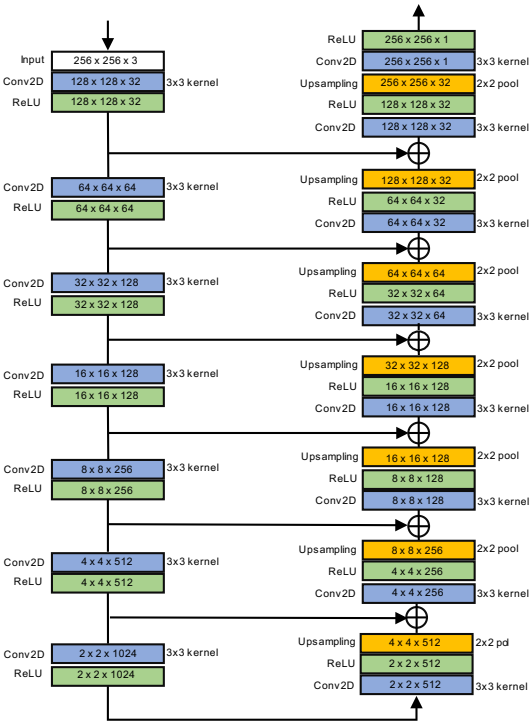
The proposed network architecture is inspired by the physical image formation process (Section 6.4), and thus the quantities involved in the rendering equation,

[1] http://rll.berkeley.edu/bigbird/
[2] https://www.shapenet.org/about
[3] http://gvv.mpi-inf.mpg.de/projects/LIME/

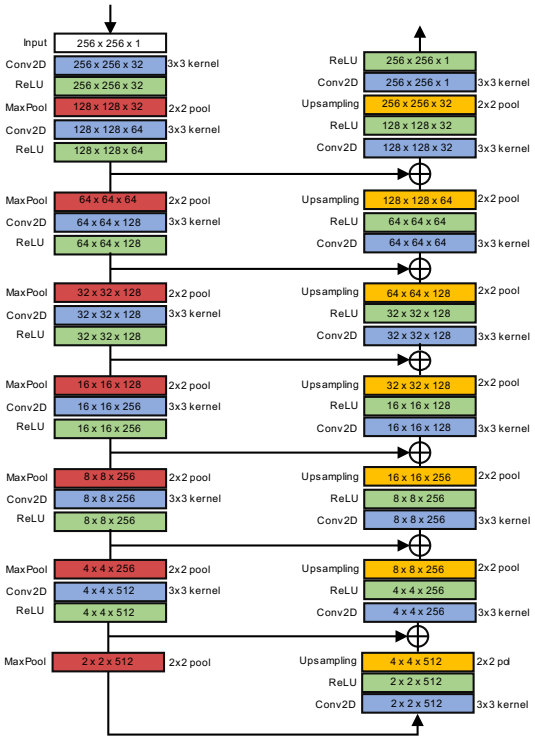
6.5 Learning



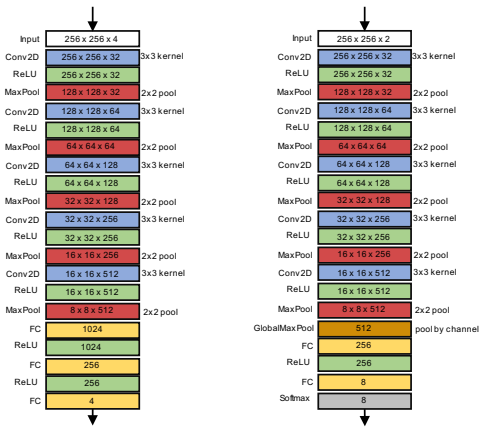
(a) The architecture of *SegmentationNet*, which learns a binary segmentation mask from a colour input image. The numbers in each box denote $width \times height \times channels$ of the layer's output, and a plus in circle represents concatenation of feature maps.



(b) The architecture of *SpecularNet*, which learns the grayscale specular decomposition from a masked colour input image.



(c) The architecture of *MirrorNet*, which learns a grayscale mirror image from a grayscale specular shading image.



(d) **Left:** The architecture of *AlbedoNet*, which learns diffuse albedo in color (3 parameters) and grayscale specular albedo (1 parameter) from the masked color input image (3 colour channels) concatenated with the grayscale specular image (1 colour channel). **Right:** The architecture of *ExponentNet*, which learns the shininess exponent using classification into 8 bins from the concatenation of the specular and mirror images (both grayscale).

Figure 6.3: Network architecture of the sub-networks of the proposed technique.

as illustrated in Figure 6.1. The task of material estimation is partitioned into five CNNs tailored to perform specific sub-tasks. The first step is the estimation of a binary segmentation mask (*SegmentationNet*) to identify the pixels that belong to the dominant object in the scene. Next, the masked input image is decomposed to obtain the specular shading image (*SpecularNet*). The mirror estimation subnetwork (*MirrorNet*) converts the specular shading image into a mirror image by removing the specular roughness. Finally, the albedo estimation network (*AlbedoNet*) uses the masked input image and the specular shading image to estimate the diffuse and specular albedo parameters. The exponent estimation network (*ExponentNet*) combines the specular shading image and the mirror image to produce the specular exponent, which ranges from diffuse to shiny.

The proposed architecture provides the opportunity for intermediate supervision using the known ground-truth quantities from the training corpus, which leads to higher-quality regression results than direct estimation with a single CNN. The proposed approach also enables the implementation of a novel perceptual rendering loss, which is discussed in Section 6.5.3. While the network is based on five sub-tasks, it is trained end-to-end, which typically results in better performance than using individually trained components. As shown in the results section, the core representation, which is based on specular and mirror images, is better suited for the image-to-image translation task than the direct regression of reflectance maps in the previous work of Georgoulis *et al.* (2017b). The main reason is that the corresponding image-to-image translation task is easier, in the sense that the CNN has only to learn a per-pixel colour function, instead of a colour transform in combination with a spatial reordering of the pixel, as is the case for reflectance and environment maps. This is because the pixel locations in reflectance and environment maps inherently depend on the underlying unknown scene geometry of the real world object. The estimated mirror image, in combination with the specular image, enables the regression of material shininess with higher accuracy, since it provides a baseline for exponent estimation.

The input to the novel inverse rendering network are 256×256 images that contain the full object at the center. The architectures of *SegmentationNet* (Figure 6.3a), *SpecularNet* (Figure 6.3b) and *MirrorNet* (Figure 6.3c) follow U-Net (Ronneberger *et al.*, 2015). The skip connections allow for high-quality image-to-image translation. *AlbedoNet* is an encoder with 5 convolution layers, each followed by ReLU and max-pooling, and 3 fully-connected layers, see (Figure 6.3d). *ExponentNet* is a classification network that uses a one-hot encoding of the eight possible classes of object shininess. Binned shininess classes are used to represent just-noticeably different shininess levels, as regression of scalar (log) shininess exhibited bias towards

shiny materials (see Section 6.7 for discussion). During training, an ℓ_2-loss is applied with respect to the ground-truth on all intermediate physical quantities and the output material parameters, except for *SpecularNet* and *MirrorNet*, for which an ℓ_1-loss is used to achieve high-frequency results, and a cross-entropy loss for classification of shininess using *ExponentNet*. In addition, to further improve decomposition results, a novel perceptual rendering loss, which is described next, is applied.

6.5.3 Perceptual Rendering Loss

Since the proposed approach is trained on a synthetic training corpus (see Figure 6.2), ground-truth annotations are available for all involved physical quantities, including the ground-truth BlinnPhong parameters $\mathbf{BP}_i$. One straightforward way of defining a loss function for the material parameters is directly in parameter space, e.g., using an ℓ_2-loss. It is shown that this alone is not sufficient, since it is unclear how to optimally distribute the error between the different parameter dimensions, such that the parameter error matches the perceptual per-pixel distance between the ground-truth and the corresponding re-rendering of the object. Another substantial drawback of imposing a loss individually on each parameter is that the regression results are not necessarily consistent, i.e., the re-rendering of the object based on the regressed parameters perceptually may not match the input image, since errors in the independent components accumulate. To alleviate these two substantial problems, an additional perceptual rendering loss is proposed that leads to results of higher quality. The effectiveness of this additional constraint is shown in Section 6.7.

The novel perceptual loss is based on rewriting the rendering equation (Equation (2.1)) in terms of the diffuse shading $\mathbf{D}$ and the specular shading $\mathbf{S}$:

$$\begin{aligned}\mathbf{L}(\mathbf{x},\boldsymbol{\omega}_o) = \mathbf{m}_a + \mathbf{m}_d \underbrace{\int_{\Omega} (\boldsymbol{\omega}_i \cdot \mathbf{n}) \mathbf{E}(\boldsymbol{\omega}_i)\, d\boldsymbol{\omega}_i}_{\mathbf{D}} \\ + \mathbf{m}_s \underbrace{\int_{\Omega} (\mathbf{h} \cdot \mathbf{n})^s \mathbf{E}(\boldsymbol{\omega}_i)\, d\boldsymbol{\omega}_i}_{\mathbf{S}} \\ = \mathbf{m}_a + \mathbf{m}_d \mathbf{D} + \mathbf{m}_s \mathbf{S}. \end{aligned} \tag{6.3}$$

For high efficiency during training, the diffuse and specular shading integrals are pre-computed per pixel in the ground-truth training corpus, and stored in the form of diffuse shading and specular shading maps $\mathbf{D}$ and $\mathbf{S}$, respectively.

The perceptual rendering loss $\mathbf{R}$ directly measures the distance between the rendered prediction and the input image $\mathbf{I}$:

$$\mathbf{R}(\hat{\mathbf{m}}_d,\hat{\mathbf{m}}_s,\hat{\mathbf{S}}) = \left\| \mathbf{B} \odot \left[\mathbf{I} - \underbrace{(\mathbf{A} + \hat{\mathbf{m}}_d \mathbf{D} + \hat{\mathbf{m}}_s \hat{\mathbf{S}})}_{\text{rendered prediction}} \right] \right\|. \tag{6.4}$$

Here, $\mathbf{B}$ is the binary foreground mask, and the rendered prediction is based on the ambient colour $\mathbf{A} = m_a \mathbf{1}_3$, the predicted diffuse albedo $\hat{\mathbf{m}}_d$, the ground-truth diffuse shading $\mathbf{D}$, the predicted specular albedo $\hat{\mathbf{m}}_s$ and specular shading $\hat{\mathbf{S}}$. The specular shading $\hat{\mathbf{S}}$ is directly predicted instead of the shininess s to alleviate the costly integration step over the environment map. Since all physical quantities are pre-computed in the training corpus, the rendering step is a simple per-pixel operation that is highly efficient and can be implemented using off-the-shelf operations such as per-pixel addition and multiplication, which are already provided by deep-learning libraries, without the need for a hand-crafted differentiable rendering engine as in Liu *et al.* (2017).

6.5.4 End-to-End Training

All the networks are trained using TensorFlow (Abadi *et al.*, 2015) with Keras (Chollet *et al.*, 2015). For fast convergence, the novel inverse rendering network is trained in two stages: First, all subnetworks are trained separately based on the synthetic ground-truth training corpus. Then *SpecularNet* and *MirrorNet* are trained together. Afterwards, *ExponentNet* is added, and finally the *AlbedoNet* is added to the end-to-end training. The gradients are back-propagated to update the network parameters using Adam (Kingma & Ba, 2015) with default parameters. First the networks are trained for 100,000 iterations with a batch size of 32, and then fine-tuned end-to-end for 45,000 iterations, with a base learning rate of 0.0001 and $\delta = 10^{-6}$.

6.6 Temporal Fusion

The single-shot inverse rendering approach estimates plausible material parameters from a single image. However, when applied to video streams independently per video frame, the estimation may have some temporal instability due to changing lighting conditions, camera parameters or imaging noise. To improve the accuracy and temporal stability of the approach, it is therefore proposed to temporally fuse all the estimated parameters. This leads to results of higher quality and higher temporal stability. A sliding window median filter with a window size of 5 frames is

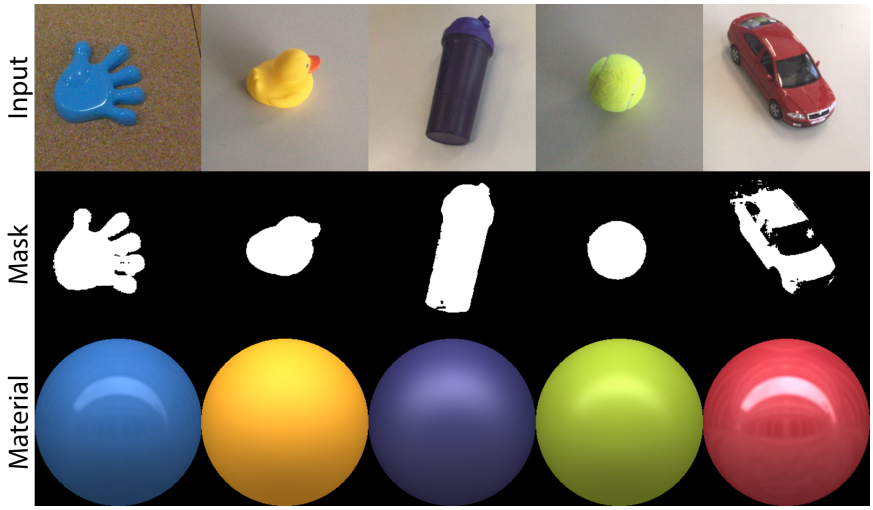


Figure 6.4: Real world material estimation results based on a single colour image. The proposed approach produces high-quality results for a large variety of objects and materials, without manual interaction.

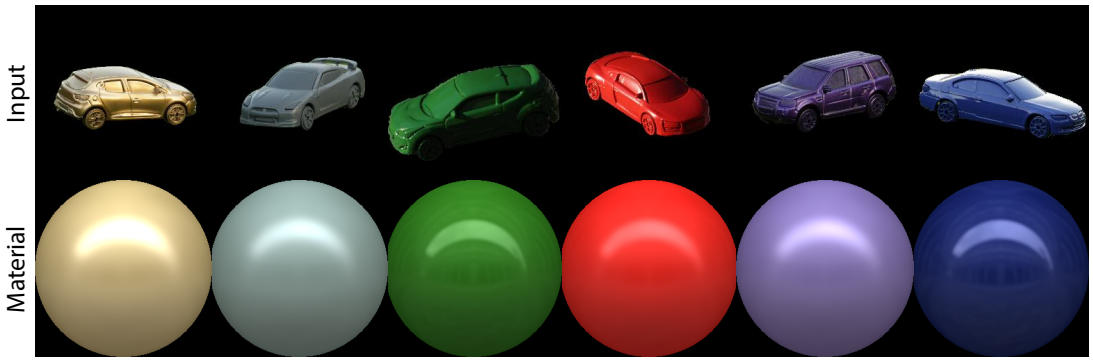


Figure 6.5: Material estimates on the dataset of Rematas *et al.* (2016).

used. This helps to filter out occasional outliers. From the median-filtered output, a decaying exponential averaging is performed:

$$\mathbf{P}^t = \alpha \widehat{\mathbf{P}} + (1-\alpha)\mathbf{P}^{t-1}. \tag{6.5}$$

Here, $\widehat{\mathbf{P}}$ is the current parameter estimate, $\mathbf{P}^t$ is the final estimate for the current time step t, and $\mathbf{P}^{t-1}$ is the fused result of the previous frame. A decaying blending factor $\alpha = (1/t)$ is used for all the experiments. This temporal filtering and fusion approach is particularly useful for the environment map estimation strategy (see Section 6.8), since it helps in integrating novel lighting directions sampled by the object as the camera pans during the video capture.

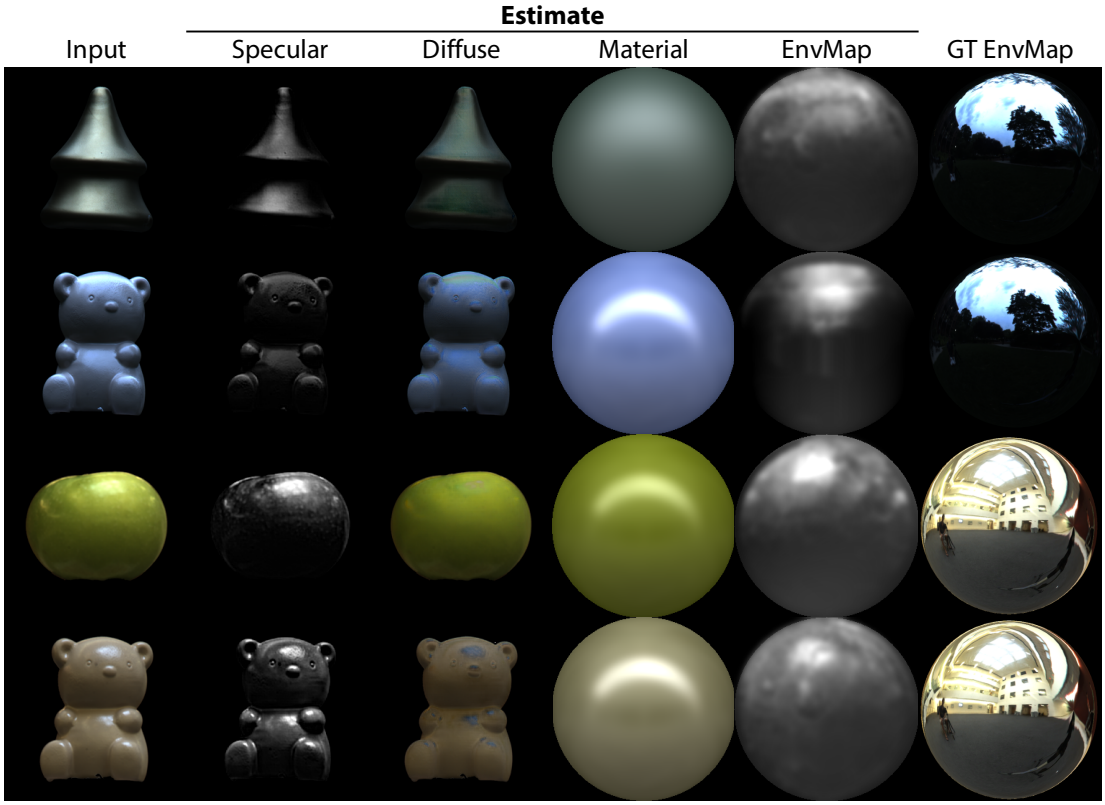


Figure 6.6: The proposed approach estimates the specular decomposition layer from a single colour image. The diffuse layer can be obtained by subtracting it from the input. With the available ground-truth normals, the environment map is reconstructed using the technique described in Section 6.8. The input images are from Lombardi & Nishino (2016b).

6.7 Experiments

Qualitative and quantitative results of the real-time single-shot material estimation approach are shown, compared to state-of-the-art approaches, and finally the design decisions are evaluated to show the benefits of the novel perceptual loss and physically-motivated network architecture.

6.7.1 Qualitative Results

Figure 6.4 shows real-time material estimation results for a wide range of different materials and general objects. As can be seen, the proposed approach estimates material parameters at high-quality for many challenging real world objects that have uniform material, without the need for manual interaction.

	Shininess Exponent (correct bin + adjacent bins)	Average Error		
		Diffuse Albedo	**Specular Albedo**	**Shininess ($\log_{10}$)**
Our full approach	45.07% + 40.12%	0.0674	0.2158	0.3073
without perceptual loss (Section 6.5.3)	45.15% + 40.96%	0.1406	0.2368	0.3038
without MirrorNet	36.29% + 40.28%	0.0759	0.2449	0.3913
with exponent regression (log10)	44.09% + 41.28%	0.0683	0.2723	0.2974
Reflectance Map Based Estimation	13.57% + 25.29%	0.0408	0.1758	0.7243

Table 6.1: Quantitative evaluation on a test set of 4,990 synthetic images. The column Shininess Exponent shows the accuracy of exponent classification, reported as percentage classified in the correct bin and the adjacent bins. The last three columns show the direct parameter estimation mean square error over the full test set. Please note that the error on shininess is evaluated in log-space to compensate for the exponential bias.

The proposed approach was also applied to the photos of painted toy cars by Rematas *et al.* (2016), shown in Figure 6.5, and high-quality material estimates were obtained. In addition, the approach can estimate the specular shading layer from a single colour image, which enables us to compute the diffuse shading layer by subtraction, as shown in Figure 6.6. Note that the approach works for general objects, and does not require manual segmentation. In contrast, previous techniques such as Georgoulis *et al.* (2017b); Liu *et al.* (2017); Lombardi & Nishino (2016b) either work only for a specific object class or require known segmentation.

6.7.2 Run-time Performance

On an Nvidia Titan Xp, a forward pass of the complete inverse rendering network takes 13.72 ms, which enables various live applications discussed in Section 6.8. The individual run times are: *SegmentationNet* (2.83 ms), *SpecularNet* (3.30 ms), *MirrorNet* (2.99 ms), *AlbedoNet* (2.68 ms) and *ExponentNet* (1.92 ms).

6.7.3 Quantitative Evaluation and Ablation Study

The method's performance is quantitatively analysed to validate the design choices. Average estimation errors are compared for groups of material parameters on an unseen test set of 4,990 synthetic images in Table 6.1. The full proposed approach is compared to three alternative versions by modifying one aspect of the network in each instance, plus one alternative:

1. The proposed network as-is, but without the novel perceptual loss (Section 6.5.3). Exclusion of the perceptual loss leads to reduced accuracy in the albedo estimates.

Actual \ Predicted	1	2	3	4	5	6	7	8
1	56%	30%	6%	3%	5%	0%	0%	0%
2	19%	51%	23%	5%	2%	0%	0%	0%
3	9%	24%	42%	20%	4%	1%	1%	0%
4	8%	11%	28%	29%	21%	3%	1%	0%
5	7%	2%	8%	23%	44%	12%	4%	0%
6	2%	1%	2%	7%	37%	35%	15%	1%
7	1%	0%	2%	3%	8%	25%	49%	12%
8	0%	0%	0%	0%	4%	5%	34%	55%

Actual \ Predicted	1	2	3	4	5	6	7	8
1	48%	32%	10%	5%	3%	1%	1%	0%
2	21%	47%	25%	6%	1%	0%	0%	0%
3	6%	25%	41%	21%	5%	1%	0%	0%
4	6%	9%	25%	36%	19%	4%	1%	0%
5	3%	4%	8%	24%	43%	11%	6%	0%
6	1%	1%	2%	9%	36%	25%	24%	2%
7	0%	0%	1%	2%	8%	15%	49%	23%
8	0%	0%	0%	0%	2%	4%	30%	62%

Figure 6.7: Confusion matrix of shininess prediction for classification (left) and regression of log-shininess (right).

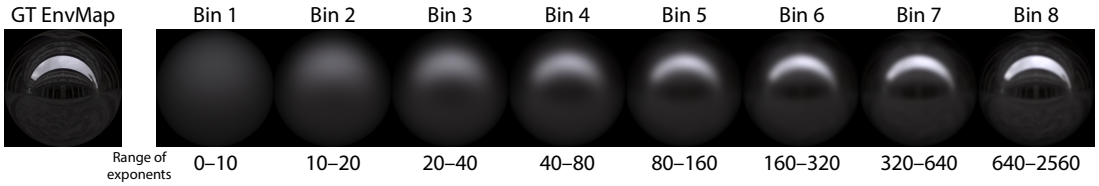


Figure 6.8: The 8 bins used for shininess exponent estimation, ranging from most diffuse (bin 1) to most shiny (bin 8). The visualization uses a material with a diffuse albedo of zero, a specular albedo of one, and shininess set to the mean value of each exponent bin. The materials are shown under the 'Uffizi' environment map Debevec (1998).

2. The proposed network without the *MirrorNet*, so that the *ExponentNet* only depends on the output of *SpecularNet*. The exclusion of *MirrorNet* leads to reduced exponent classification accuracy, thus proving the efficacy of the mirror-representation-based design on the challenging task of estimating the non-linear material shininess.

3. The network with the *ExponentNet* modified to regress shininess directly instead of as a classification task. The regression is performed in log space (base 10). The average errors show similar performance in both the original classification and this regression case. Yet, classification is chosen as the final design of the method. This choice is made because the regression network exhibits a bias towards specular materials, i.e., it performs well for specular materials, but quite poorly on diffuse materials. This becomes more evident when looking at the distribution of the estimation accuracy for shininess over the classification bins in the confusion matrix in Figure 6.7. The bins are numbered in an increasing order of log-shininess.

The range of exponent values designated to each bin is detailed in Figure 6.8. The confusion matrix for the classification task (left) is symmetric at the diffuse and specular ends, whereas for the regression (right) it is more asymmetric and biased towards specular predictions. This bias is also visible on real world data, in which case the regression network performs poorly for diffuse objects. This bias appears to result from the different losses employed in the training[1].

4. This method uses the encoderdecoder structure of Georgoulis *et al.* (2017b) to take a segmented input (from the *SegmentationNet*), and estimates a spherical reflectance map of size 256×256. This reflectance map is then fed to a second network that estimates the material parameters. Both networks are first trained independently, and then tuned by training end-to-end for a fair comparison. This approach attains slightly lower albedo errors, but performs poorly on shininess estimation. This might be due to the non-linear re-ordering required to convert an image of an object into a reflectance map, which results in the loss of high-frequency information that is essential to exponent estimation.

All networks are trained on the full training data, until convergence. The accuracy of the *SegmentationNet* is also reported on this test set as 99.83% (Intersection over Union).

6.7.4 Comparison to the state-of-the-art

The proposed approach is compared to the state-of-the-art in learning-based material estimation. First, the material transfer results are compared to Liu *et al.* (2017) and Lombardi & Nishino (2016b) in Figure 6.9. The approach of Liu *et al.* (2017) requires optimization as a post-process to obtain results of the shown quality, while the proposed approach requires just a single forward pass of the network. Here, the environment map of the target object is also computed using the intermediate intrinsic layers regressed by the proposed network (see Section 6.8). The proposed approach obtains more realistic material estimation and therefore better transfer results.

The proposed approach is also compared to the material estimation results of Georgoulis *et al.* (2017b) in Figure 6.10. The proposed method is able to estimate the colour and specularity of the cars more accurately.

[1]The classification task uses a binary cross-entropy loss which treats each bin as equal, whereas the regression task uses the mean absolute error, which may have greater error for larger exponent values, hence biasing.

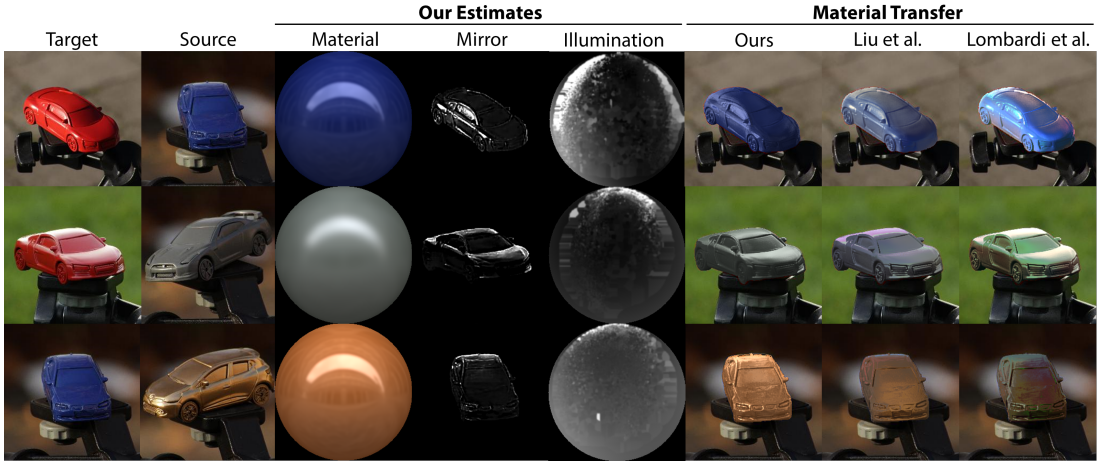


Figure 6.9: Material estimation and transfer comparison. From left to right: Image of the target object, source material to copy, the estimates of material, mirror image and environment map, and the transfer results of the proposed approach, Liu *et al.* (2017) and Lombardi & Nishino (2016b). The proposed method obtains better material estimates (top two rows) and illumination (third row). For fairness and comparability of the results, the normal map estimated by Liu *et al.* (2017) is used for the environment map estimation in this case.

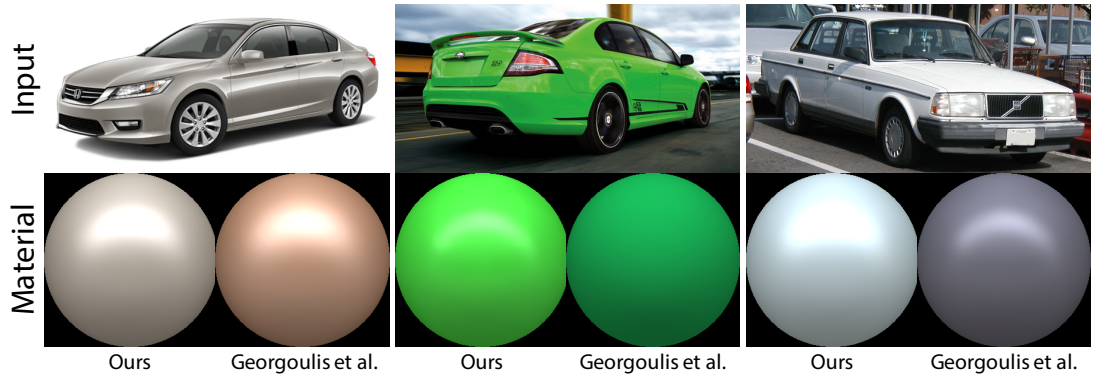


Figure 6.10: Comparison to the approach of Georgoulis *et al.* (2017b). Note that their approach is specifically trained for the outdoor estimation scenario, while the proposed approach is trained for the indoor setting. Nonetheless, this approach obtains results of similar or higher quality.

6.8 Applications

Real-time material estimation from a single image or video can provide the foundation for the following mixed- and augmented reality applications:

Single-Shot Live Material Estimation The proposed approach can estimate material parameters in a live setting, so that material properties of real world objects

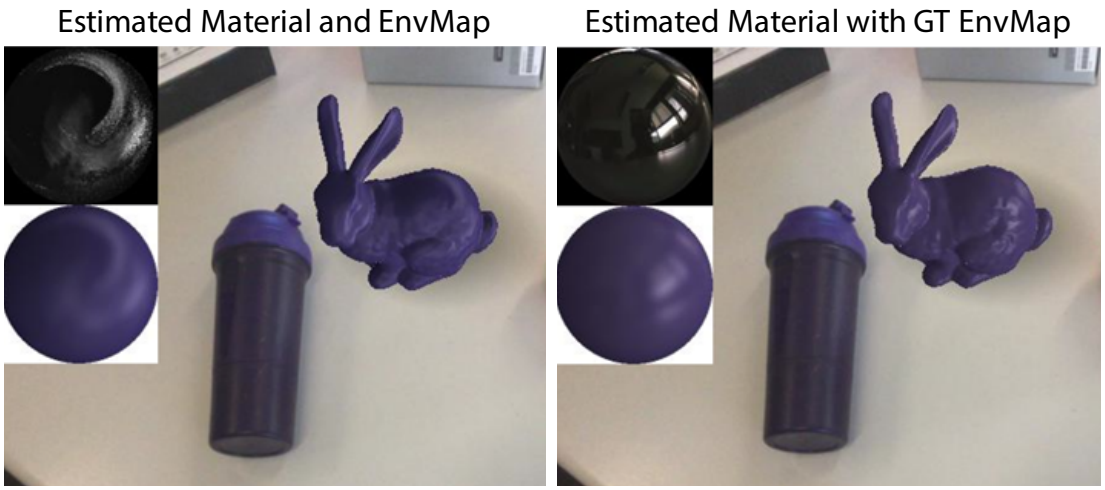


Figure 6.11: Cloning of real world materials on virtual objects in an illumination-consistent fashion. **Left:** Estimated material rendered with the estimated environment map. **Right:** Estimated material rendered with the ground-truth environment map.

can be reproduced in virtual environments from just a single image, for instance in a video game or in a VR application.

Live Material Cloning When surface geometry is available, e.g., when using a depth sensor, the proposed approach can also be extended to compute an environment map alongside material parameters. This essentially converts an arbitrary real world object into a light probe. The normals available from a depth sensor are used to map the estimated mirror image to a spherical environment map; this provides the high-frequency lighting information of the scene. The diffuse shading image of the object is also obtained and a low-frequency spherical harmonics lighting (discussed earlier in Section 2.3) estimate computed for the scene using the available normals. The full environment map lighting is obtained by adding the two. This process is followed for single image when normals are available, for example for the target image in Figure 6.9. In case of a video as input, the low- and high-frequency lighting estimates of multiple time steps are integrated into the same environment map using the filtering and fusion technique described in Section 6.6. The camera is also tracked using the real-time volumetric VoxelHashing framework of Niener *et al.* (2013) similar to Chapter 4, so that environment maps can be integrated consistently in scene space rather than relative to the camera. The estimate is then transferred to the virtual object of choice, and rendered seamlessly into the real world scene, as shown in Figure 6.11.

6.9 Discussion

The proposed approach works well for many everyday objects, but does not handle more complex BRDFs well. Particularly difficult are scenarios that violate the assumptions that the model makes. The proposed approach may fail in the presence of global illumination effects such as strong shadows or inter-reflections. While most commonplace dielectric materials exhibit white specularity, some metallic objects have coloured specularity, which the current approach does not support. This could be addressed with more expressive BRDF and global illumination models. The quality of the material and environment map estimates depends on the quality of the input data. Modern cameras provide good white balancing, and the white illumination model hence fits well for many indoor scenarios, yet some special lighting arrangements, such as decorative lighting, require handling of colour illumination. Working with low-dynamic-range images also implies dealing with camera non-linearities, which may lead to saturation artifacts, e.g., in the teddy bear in Figure 6.6. The quality of surface normals derived from depth sensors is currently not adequate for accurate high-frequency illumination estimation. Future AR and VR devices with more advanced depth sensing capabilities may help to improve the quality of estimated environment maps.

6.10 Conclusion

The first real-time approach for estimation of diffuse and specular material appearance from a single colour image is presented. The highly complex and ill-posed inverse rendering problem is tackled using a discriminative approach based on image-to-image translation using deep encoderdecoder architectures. The proposed approach obtains high-quality material estimates at real-time frame rates, which enables new mixed-reality applications, such as illumination-consistent insertion of virtual objects and live material cloning.

While this chapter looks at acquiring the reflectance of a uniform material object, the next chapter (Chapter 7) looks at the more complex case of acquiring spatially varying reflectance in real-time. This is a severely under-constrained problem which is dealt with using the specialized hardware setup of a Light Stage, in particular for the one very important class of everyday objects – human faces.

6.10 Conclusion

Chapter 7

Deep Reflectance Fields

This final technical chapter of the thesis looks at the very challenging problem of estimating spatially varying reflectance fields in real-time, particularly for the case of dynamic human faces. Human faces are of interest to several augmented and virtual reality use-cases such as tele-presence and acquiring the full reflectance field of facial performances enables photorealisitc relighting, and hence very immersive graphical experiences. Since this is a dense and highly under-constrained problem, specialized hardware in the form of a Light Stage is used to acquire training data for a novel discriminative approach. The resulting technique (Meka *et al.*, 2019a) acquires at frame rate a deep reflectance field of human faces including all light transport effects such as shadows, shading, specularities and sub-surface scattering, while generalizing accurately across viewpoint, identity and expression.

7.1 Introduction

Modifying the lighting in a facial portrait image is a much sought after capability that would benefit many visual effects including portrait photography and virtual or augmented reality applications.

This relighting is particularly challenging since the facial appearance is the result of a complex interaction of light with the many materials that make up the skin, eyes, hair, teeth, and clothing, each of which have complex geometry and varying amounts of specular reflection and subsurface scattering. Further, ignoring or approximating these properties is especially perilous as humans are highly capable of detecting the subtle cues of realism in facial renderings. While today's computer graphics techniques can produce photorealistic digital human models which can be rendered in any lighting and from any viewpoint, creating such models is still

extremely laborious and expensive. Indeed, progress towards automated avatar creation still falls far short of photorealism.

In order to reach the highest level of photorealism, image-based relighting systems capture actors at a high resolution under a large number of lighting conditions. For instance, high-quality pore-level 4D reflectance fields of humans can be acquired with the Light Stage proposed by Debevec *et al.* (2000) – a spherical dome equipped with a large number of controllable light sources and cameras. The 4D reflectance field from one camera view can be sampled by capturing hundreds of one-light-at-a-time (OLAT) images, each of them capturing the subject illuminated by a single light on the Light Stage. By projecting the environment map of a new illumination condition onto this captured illumination basis, photorealistically re-lit images of a subject can be created as a weighted combination of the OLAT images. The relighting results exhibit the full range of local and global effects, including diffuse lighting, specular reflections, inter-reflection, subsurface scattering, and self-shadowing. Unfortunately, capturing several hundreds OLAT images, a number typically required for high-quality reflectance field capture, requires several seconds, e.g., $\approx$8 seconds using the Light Stage 2 of Debevec (2012). Capturing a time-varying reflectance field of dynamic scenes in this way is challenging, and relies on a hardware setup variant equipped with high speed cameras, as well as an error-prone optical flow alignment step as seen in Einarsson *et al.* (2006).

To allow the capture of dynamic scenes, the key is to be able to rely on a *small* set of input images that can be captured at real-time frame rates – while the actor is performing freely. In this setting, strong priors can help to better constrain reconstruction, but they introduce significant trade-offs. For instance, Saito *et al.* (2017) and Yamaguchi *et al.* (2018) only handle skin, and can not correctly relight facial hair, eyes, teeth, accessories, or upper body clothing, since their underlying assumptions do not hold in these regions. An alternative to the manually crafted priors is the use of learnable pipelines such as the one proposed by Xu *et al.* (2018). Their deep neural network seeks to relight a scene under novel illumination based on a set of five optimal images captured under predefined directional lighting. The approach provides compelling results on synthetic data, but fails to handle complex object shapes and high-frequency details such as shadows, and can only handle low image resolutions (128×128 pixels).

In this chapter, a new approach for the acquisition of high-quality time-varying 4D reflectance fields of a human actor at 30 fps in a Light Stage is introduced, without having to resort to time-multiplexing, motion compensation techniques, or priors. The proposed approach uses a deep neural network to learn a mapping from only *two* images, captured under spherical gradient illumination, to the full 4D reflectance

field. As such, it can reconstruct *any* OLAT image from a given lighting direction. The predicted dynamic reflectance fields come very close in quality to the reflectance models captured with a dense set of OLAT images. The method enables quasi-photorealistic relighting of the *complete* human head as it handles skin subsurface scattering, wrinkle details, skin specularity, facial hair, and teeth, as well as the complex appearance of the human eyes in a unified manner, and in a way that generalizes across different identities. While a Light Stage only generates a *discrete* illumination basis due to the finite number of mounted light sources, the proposed method recovers a *continuous* illumination basis, since the network can be evaluated for any illumination direction. The core technical contributions can be summarized as:

- A capture system that enables 4D reflectance field estimation of moving subjects.
- A machine learning-based formulation that maps spherical gradient images to the OLAT image corresponding to a particular lighting direction.
- A task-specific perceptual loss trained to pick up specularities and high-frequency details.
- An alignment loss that robustly handles the small misalignments between the spherical gradient images and the ground-truth OLAT images.

Experiments show that the method is effective in real applications such as relighting in arbitrary lighting environments and compares favourably with offline capture systems and other state-of-the-art approaches.

7.2 Related Work

Modelling photorealistic humans is an active research topic in the computer vision, graphics, and machine learning communities. Related works that are representative of different trends in the literature as *parametric model fitting*, *image-based*, and *learning-based* solutions are discussed here.

Parametric model fitting These approaches assume strong priors, typically performing an explicit reconstruction while employing hand-designed reflectance and/or lighting models. General shape, illumination, and reflectance can be recovered based on a set of hand-crafted priors and optimization as described in Chapter 4 and other state-of-the-art approaches such as that of Barron & Malik (2015b). Parametric models of geometry, surface reflectance, or illumination have been employed for reconstruction and relighting in the context of human bodies in Theobalt *et al.*

(2007), faces in Blanz & Vetter (1999); Garrido *et al.* (2013, 2016); Gotardo *et al.* (2018); Hawkins *et al.* (2004); Ichim *et al.* (2015); Thies *et al.* (2016), eyes in Bérard *et al.* (2016), eyelids in Bermano *et al.* (2015), and hair in Hu *et al.* (2015); Zhang *et al.* (2017). Faces can be relit under a diffuse appearance assumption based on radiance environment maps and ratio-images (Wen *et al.*, 2003). Other approaches jointly estimate parametric BRDF models and wavelet-based incident illumination to relight 3D videos of humans (Li *et al.*, 2013). Relighting of the human head can be formulated as a mass transport problem as shown by Shu *et al.* (2017) based on position and normal estimates recovered by a parametric face model. Cosine lobe relighting can be performed analytically based on a pair of spherical gradient illumination images (Fyffe *et al.*, 2009), but secondary effects such as shadows are of low quality due to the use of approximations in modelling the face geometry. Some recent deep learning-based approaches such as Saito *et al.* (2017); Yamaguchi *et al.* (2018) estimate the *parameters* of a predefined reflectance model from single images. The approach of Gotardo *et al.* (2018) for dynamic appearance estimation extracts sv-BRDF (diffuse and specular) and geometry (also fine scale) from images captured under uniform lighting, but their approach is restricted to the skin region. Recently, multiple works have also focused on the challenging problem of extracting the sv-BRDF from a *single* image using a flash (Li *et al.*, 2018a,b; Nam *et al.*, 2018). Since all model-based approaches use hand-crafted priors, they are typically limited to specific parts of the human body and only handle these in isolation. Many of these approaches only work under low-frequency illumination conditions and do not handle the specularity of skin and sub-surface scattering effects. In contrast, the proposed model-free approach enables relighting of the *complete* human head.

Image-based relighting To reach the highest level of realism, image-based relighting techniques capture actors at a high resolution under a large number of lighting conditions. High-quality pore-resolution 4D reflectance fields of humans can be acquired with a Light Stage as shown by Debevec *et al.* (2000). Einarsson *et al.* (2006) illuminates the scene with a smaller set of approximately 30 lighting basis functions with larger spatial support to enable real-time capture, but this comes at the expense of lighting resolution. Other techniques such as Wenger *et al.* (2005) use high frame rate video and time-multiplex the sampling of the lighting basis over a window of several frames, but this requires expensive and error prone motion estimation. An alternative approach is to use a reference subject's 4D reflectance field to modify the lighting on a target subject's performance using an aligned ratio image as done by Peers *et al.* (2007). However, this requires having a 4D reflectance field available of

a similar-looking subject and can transfer high-frequency details from the reference subject to the target. For dynamic performances, this solution is approximate as it interpolates from a sparsely sampled collection of static poses. The style transfer technique of Shih *et al.* (2014) matches local image statistics from a reference portrait to a target portrait and thereby is also able to perform some degree of relighting of the target portrait. However, the techniques require manual touch-up and can be challenged by harsh lighting scenarios. Unfortunately, the acquisition of 4D reflectance fields is a slow process and thus the subject would have to move in a stop-motion manner. This makes capturing high-quality reflectance fields of dynamic facial performances very difficult, requiring expensive high speed cameras running at thousands of frames per second and potentially uncomfortable light levels (Wenger *et al.*, 2005). To the best of my knowledge, the proposed method introduces the first approach for deriving *time-varying* 4D reflectance fields of a human actor at 30 fps in a Light Stage.

Learning-based techniques Deep learning based techniques have recently been applied to the problem of relighting arbitrary objects as show in the previous chapter and other techniques (Chapter 6, Meka *et al.* (2018); Ren *et al.* (2015); Xu *et al.* (2018)) and human bodies (Kanamori & Endo, 2018). The method of Nalbach *et al.* (2017) showed that appearance synthesis can be cast as a learning based screen-space shading problem based on per-pixel scene attributes such as position, normal and reflectance. Based on a set of OLAT images, the approach of Xu *et al.* (2018) is trained to relight a scene under novel illumination based on an optimal set of five jointly selected OLAT images. While results are compelling, it fails to handle complex object shapes, high-frequency specularities, and shadows caused by grazing angle illumination and non-convex geometry. The data-driven rendering of Lombardi *et al.* (2018) learns a joint representation of facial geometry and appearance from a multi-view capture setup, but this technique does not address the problem of relighting. The approach by Kanamori & Endo (2018) enables occlusion-aware inverse rendering for the human body, but results are restricted to Lambertian surfaces and low-frequency illumination. In contrast, a novel machine learning-based formulation is proposed that maps spherical gradient images to a full dataset of one-light-at-a-time (OLAT) images. This enables *model-free* relighting of dynamic scenes captured in a Light Stage.

In contrast to all other approaches, this method leverages the insight of Fyffe

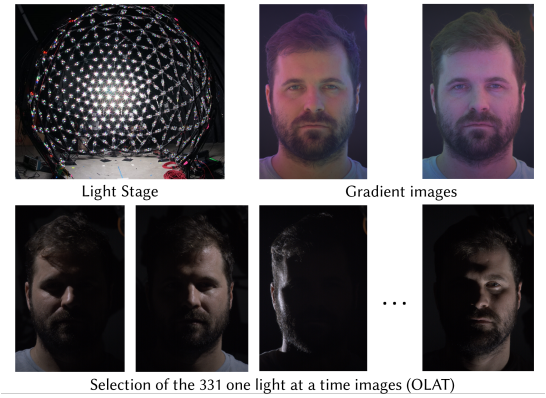


Figure 7.1: **Capture setup** – Programmable light sources mounted on a geodesic dome are used to light the subject under RGB colour gradient images and OLAT images for training, inference, and validation data.

et al. (2009), where spherical gradient images are used to derive full-colour diffuse/specular albedo and surface normals for dielectric materials. Fyffe *et al.* (2009) assumed a simple cosine lobe model, which uses only *local* information and low frequency statistics to fit the BRDF. It is shown here that, if a more expressive underlying model is provided, spherical gradient images contain all the information necessary to generate the full reflectance field. In the proposed method, this model is a neural network that infers the complex mapping from spherical gradient images to every possible directional lighting condition. The model can take advantage of *non-local* information and contextual cues. For the first time, this enables estimating full reflectance fields of dynamic subjects without any explicit prior or BRDF model.

7.3 Capture Setup

As light follows the superposition principle, one can photorealistically apply any desired lighting configuration to a given actor by combining a finite set of lighting conditions. In more detail, by capturing a set of images where only one light is turned on at a time (OLAT) using the capture setup shown in Figure 7.1, one can linearly combine the RGB channels of these images in order to simulate a desired environment map; see Figure 7.2. In practice, sufficiently high sampling resolution in both captured images and light sources is key to ensure that details in both the surface (e.g., skin pores) and directional effects (e.g., specularities) are captured. The main disadvantage of this approach is the extended duration of time during which the subject has to remain still while the OLAT images are captured. As there are 331 lights in the system, the acquisition of the corresponding 331 OLAT images would take several seconds, making the capture and relighting of dynamic performances a real challenge. One of the contributions of this work is overcoming

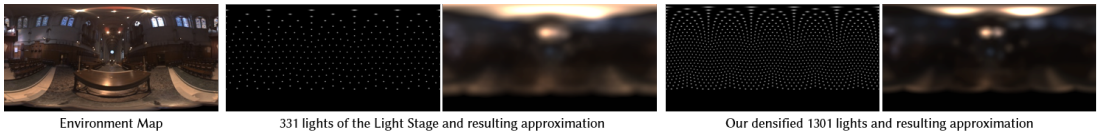


Figure 7.2: **Image-Based Relighting** – An environment map (left) can be approximated with the 331 lighting directions of the Light Stage (middle). With a denser sampling of 1301 light directions, as enabled by the proposed method, it is possible to obtain a better lighting environment approximation (right).

this limitation by directly regressing an arbitrary OLAT image using *only two* observations of the subject captured under spherical gradient illumination.

7.3.1 Spherical colour gradient images

Spherical colour gradient illumination images for reflectance estimation were originally proposed in Fyffe *et al.* (2009). Given the lighting direction vector θ of a LED relative to the centre of the Light Stage, the light emitted by that LED for the first gradient image is programmed to have the RGB colour $((1+\theta_x)/2,(1+\theta_y)/2,(1+\theta_z)/2)$, and the second gradient image is programmed to have the RGB colour $((1-\theta_x)/2,(1-\theta_y)/2,(1-\theta_z)/2)$. Figure 7.1 shows the two gradient images captured from a single camera viewpoint. Although simple to form, these images can be leveraged to recover important reflectance information about the surface being captured as shown by Fyffe & Debevec (2015); Fyffe *et al.* (2009). In particular, the patterns, when summed, produce a full-on white light condition which reveals the subject's *total reflectance* (diffuse plus specular), and the difference of the images encodes the average reflectance direction into the RGB colour channels (a strong cue for surface normals). Further, the magnitude of the difference image relative to the sum image is a function of not only the BRDF but also the local self-shadowing (cues to shadow estimation). In this sense, the photographs under the two illumination patterns provide both geometric and albedo information to the inference algorithm. In contrast to previous work that interpreted gradient images using simple local parametric reflectance models, this approach employs deep learning to leverage the spatial context of the gradient images to infer far more realistic reflectance estimates.

7.3.2 Hardware and data capture

To acquire the necessary spherical gradient observations with corresponding ground-truth OLAT images for training, an LED sphere Light Stage capture setup (Debevec,

2012) is leveraged. The Light Stage is a 3.5m diameter spherical dome on which 331 custom LED light sources with red, green, blue, and white controllable LEDs are evenly distributed as in Figure 7.1. Each of these LEDs is fully controllable by a driver, allowing it to emit light of any desired intensity and colour. In order to capture actors at high resolution and under different viewpoints, nine Sony IMX253 cameras, capable of capturing 12.4 MP images at 60 Hz are used. All of the lights and cameras are synchronized via a hardware trigger.

Data capture and post-processing As relighting is cast here as a supervised regression problem, corresponding inputs and outputs are required to train the neural network; see Figure 4.1. The input consists of two colour spherical gradient images and a desired lighting direction, while the output is an OLAT image corresponding to that lighting direction. In order to relight an image at test time,a collection of OLAT images (the full 4D reflectance field) are predicted using only the two spherical gradient images as input. Note that the OLAT images are only captured for *training* purposes and thus, at *inference* time, only gradient images are captured for the dynamic sequences to be relit. Precise pixel-to-pixel correspondence between OLATs and gradient images at training time is crucial to infer sharp OLAT images at inference time. Unfortunately, it is challenging for actors to remain completely still for the extended amount of time required to capture all 331 OLAT images. To overcome this challenge, when capturing training data, "tracking frames" are interleaved into the capture sequence:

1. Capture the 331 OLAT images, however:

   (1.1) After every 11 OLAT captures, capture a "tracking frame"

2. Capture the two gradient images

A tracking frame is an image captured where *all* the lights on the Light Stage are turned on to generate homogeneous illumination. Once the capture session is over, the last tracking image is taken as a reference to compute a dense optical flow-field across tracking frames using the method by Anderson *et al.* (2016). The homogeneous illumination in the tracking frames is what makes the computation of dense optical flow possible. The optical flow field computed over tracking frames, is then linearly interpolated through time to provide correspondences across the OLATs. Although this procedure generally provides flow-fields of sufficiently good quality, the motion compensated frames can still present misalignments that could hinder

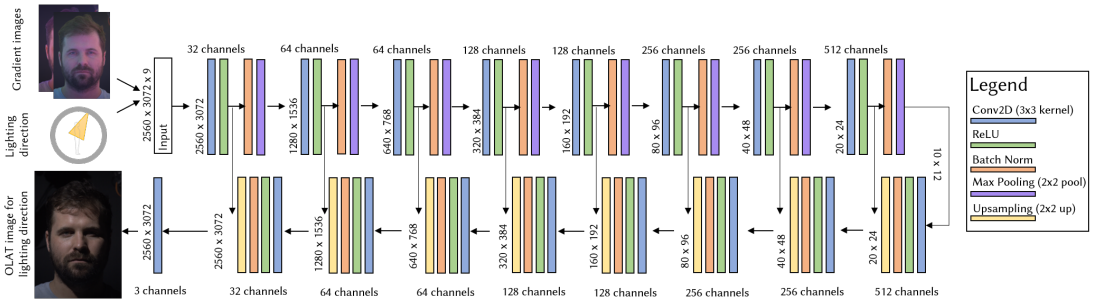


Figure 7.3: **Pipeline** – The network receives as input a pair of gradient images and a lighting direction. Via a U-Net architecture, it regresses the OLAT image that is corresponding to that particular lighting configuration.

the performance of the regressor. This issue is addressed by a novel training loss that effectively compensates for small misalignments in image space; see Section 7.4.1.

7.4 Learning

7.4.1 Predicting photorealistic 4D reflectance fields

In this section the main algorithmic contribution is described: a deep neural network capable of predicting photorealistic 4D reflectance fields for previously unseen faces. In more detail, given two gradient images and a lighting direction as input, the goal is to predict how any subject would look under white light coming from a specified spotlight direction. The OLAT prediction can be seen as solving an image-to-image translation task similar to Chen & Koltun (2017); Isola *et al.* (2017); Zhu *et al.* (2017), where the goal is to start from input images from a certain domain and "translate" them into another domain. This scenario is similar in the sense that gradient illumination images are being transformed to another image with the same content, but different illumination.

Similar to Chapter 6, the architecture that is employed is inspired by U-NET (Ronneberger *et al.*, 2015) which has recently shown impressive results on image-to-image translation tasks involving photorealistic images of humans in Martin-Brualla *et al.* (2018). A fully convolutional variant (Long *et al.*, 2015) is employed, allowing efficient training of the network on *patches* and processing of high resolution images. At inference time, the input of the network is two spherical gradient images of resolution $W = 2560$ and $H = 3072$. Similar to Eslami *et al.* (2018), the lighting direction is concatenated to *each* pixel of the input tensor. This results in an input tensor of size $W \times H \times 9$. The output of the network is an RGB image of size $W \times H \times 3$.

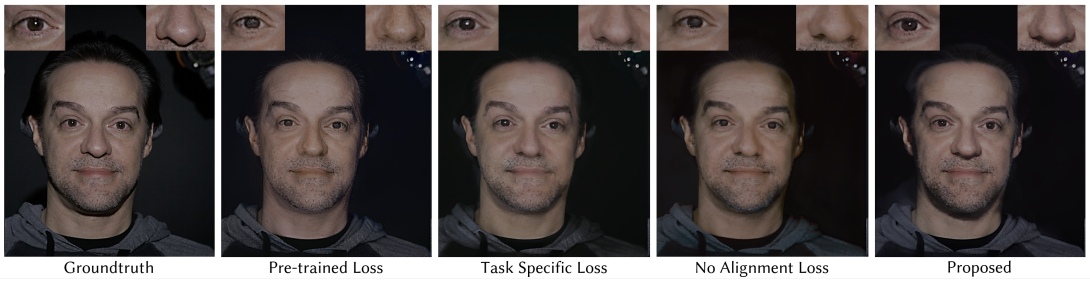


Figure 7.4: **Training losses** – Effect of different training losses on the final results. (a) ground truth, results generated with: (b) VGG pre-trained on ImageNet, (c) task dependent specific loss, (d) without alignment loss, (e) proposed loss.

The U-NET encoder takes the input tensor and runs $M = 8$ convolutional layers using 3×3 convolutions. The output of the convolutions is immediately passed through a ReLU activation function, followed by a batch-normalization layer, and a max-pool layer. In the decoder stage a bilinear upsampling is used followed by a convolutional layer. Skip connections are used between the encoder and the decoder by concatenating the output from the encoder convolutional layer to the features at the corresponding decoder layer. The network is illustrated in Figure 4.1.

7.4.2 Training

At training time random crops with resolution $512 \times 512 \times 9$ are employed as input to the network. After the $M = 8$ convolutional layers this produces a tensor of size $2 \times 2 \times 512$. Crops are crucial to train fast enough on high resolution images and to achieve the highest level of quality. They effectively limit the amount of context the network is able to see and hence prevent over-fitting (Kuo *et al.*, 2018). Using crops during training also enables the formulation of a novel patch-based *local* alignment strategy.

Training setup In order to hasten training, training is distributed across 12 NVIDIA Tesla V100 GPUs. At each training epoch, a training frame is randomly picked, a patch within that frame and one OLAT per GPU. The ADAM optimizer (Kingma & Ba, 2015) with a learning rate of 10^{-4}, and exponential decay of the learning rate with a rate of 0.1 every 10^6 iterations, is used. The network is optimized for 1 million iterations before the training converges.

Training losses Choosing the appropriate loss for a new task is non-trivial and requires systematic trial and error. For example, a simple photometric loss does not

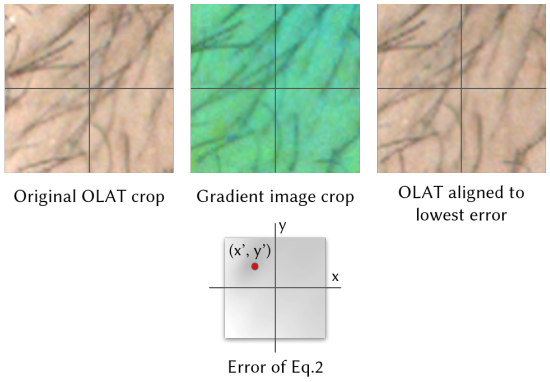


Figure 7.5: **Alignment Loss**. The slide-pooling loss accounts for misalignment between the ground truth OLAT crop (top left) and the gradient crop (top center). In the (x,y) coordinate frame (bottom) the energy of Equation 2 has a minimum at (x',y') slightly up and to the left, marked with a red dot. The ground truth OLAT image aligned to the minimum energy (top right) appears well aligned to the gradient image.

lead to photorealistic output as also shown by previous works such as Martin-Brualla *et al.* (2018). Therefore, a loss function is employed to specifically address this problem. Let I_{pred} be the prediction of the network and I_{gt} the ground-truth OLAT image, the loss is defined as:

$$\mathbf{L} = \|\text{Perc}(I_{\text{pred}}) - \text{Perc}(I_{\text{gt}})\|_2^2 \,. \tag{7.1}$$

The loss is the squared ℓ_2-norm of the difference in feature space between the predicted image and the ground-truth image. Here, the feature space is indicated by Perc$(\cdot)$. A common choice in the literature is to use a VGG network (Simonyan & Zisserman, 2014) pre-trained on ImageNet to compute the perceptual loss Zhang *et al.* (2018). While this loss is well suited for generic natural images, the task at hand is specific and such an ImageNet trained model would lead to sub-optimal results, especially when regressing specularities and other high-frequency details; see Figure 7.4. Therefore, it is proposed to enhance the loss using a VGG architecture that has been trained on a more relevant task: a random image patch sampled from a groundruth OLAT image I_{gt} is considered as input and the goal is to correctly determine which light direction generated the given patch – recall that in total there are 331 light directions. The problem is cast as a regression task and hence the networkis trained to minimize the ℓ_1-loss between the predicted direction and the ground-truth direction. Training is stable with ℓ_2 or ℓ_1 losses. An ℓ_1 loss is used as it tends to produce sharper results for image-to-image translation tasks.

As specularities heavily depend on the direction of incoming light, a perceptual loss using the new task specific VGG network is particularly sensitive to these high-frequency effects, but is inferior to a perceptual loss using a network trained on ImageNet when it comes to reconstructing lower frequencies; see Figure 7.4. As these two losses capture complementary aspects of the desired result, they are combined as $\mathbf{L} =$

$\mathbf{L}_{\text{pretrained}} + \lambda \mathbf{L}_{\text{specific}}$, where the two components $\mathbf{L}_{\text{pretrained}}$ and $\mathbf{L}_{\text{specific}}$ are obtained by using the pre-trained VGG and the task specific VGG loss respectively. In more detail, five convolutional layers from each VGG are used and the activations rescaled by their corresponding feature length to ensure that they all contribute in the same manner to the final loss. $\lambda = 0.5$ is used. The effect of this loss is shown in Figure 7.4.

Alignment loss Slight misalignments of gradient and ground-truth OLAT images leads to complications with losses that assume pixel-perfect alignment; see Section 7.3.2. Indeed, naively computing the pixel difference loss will result in blurred results. To solve this problem, a novel alignment strategy is proposed:

$$x', y' = \underset{x,y}{\operatorname{argmin}} \sum_u \sum_v \| I_{\text{gt}}(u-x, v-y) - I_{\text{pred}}(u,v) \|_1, \tag{7.2}$$

where $I(u,v)$ is the intensity value for a certain pixel location (u,v), the offsets x,y are sampled in a $[-20,20] \times [-20,20]$ window, and $\hat{x},\hat{y}$ are the optimal offsets that correspond to the best aligning image, denoted $\hat{I}_{\text{gt}}$. The image $\hat{I}_{\text{gt}}$ is then used in Equation (7.1) instead of I_{gt}, effectively producing a slide-pooling loss that takes into account translational misalignments; see Figure 7.5.

7.4.3 Inference

As described in Section 7.3.2, only two gradient images are captured per frame, allowing the capture of relightable data at 30Hz. Once the data is captured, the user only needs to define the lighting environment that should be used for relighting the captured sequences. A dense set of light directions from which to sample the environment map also has to be defined. These directions are run together with the two gradient images through the network to estimate the corresponding OLAT images. Once all the OLAT images have been obtained, they can be combined according to the environment map to form the relit images. It is interesting to note that the number of lights composing that environment map can be *much greater* than the 331 used during training, leading to more detailed relit images. For input images of 2560×3072 resolution, the inference time of the network for a single OLAT image, averaged over 100 runs, is $270.14ms$ on a workstation with an Nvidia TitanXp GPU and $1360.65ms$ on a workstation with only 2 Intel Xeon Gold 6154 CPUs. Although the inference time seems quite long, parallel GPU clusters are used to speed up the OLAT inference.

7.5 Experiments

In this section, an in-depth analysis of the proposed approach is performed. To this end, a dataset with 18 subjects is captured. For each subject, 331 ground-truth OLAT images, 2 gradient illuminations, and 33 fully lit tracking frames were captured. As mentioned in Section 7.3.2, tracking and OLAT images are only used at training time. For each person, their imagery was recorded from 9 different viewpoints. Additionally each subject was recorded giving a dynamic facial performance for 5 seconds while interleaving the two colour gradient lighting conditions. The captured data was split into a training set consisting of frames from 10 training subjects and a test set consisting of frames from 8 test subjects. Only 5 viewpoints were used for training, leaving 4 unseen viewpoints for testing.

7.5.1 Qualitative Comparisons

In this section qualitative results are shown on different test sequences and under different conditions. It is important to note that none of the subjects used for these comparisons are part of the training set.

OLAT inference In Figure 7.6, some examples of OLAT images inferred by the trained neural network are shown. The proposed method reproduces both low- and high-frequency details and achieves realistic reconstructions which closely approximate the ground-truth imagery. Shadows, reflections and details present in the original OLAT image data are faithfully reconstructed by the proposed approach.

Light direction interpolation In Figure 7.7, it is shown how the system is able to infer lighting directions that are not part of the dataset, demonstrating the method's ability to generalize. The ground-truth comprises OLAT images from 331 lighting directions. As such, a direct application of this discretized reflectance field to relight a sequence is limited in lighting resolution. For example, specular highlights that would be caused by lighting directions that are not part of the 331 sampled directions are not seen. In contrast, using the proposed network an OLAT image can be *inferred* for *any* lighting direction, essentially recovering a continuous reflectance field, as opposed to a discretized version obtained by the Light Stage.

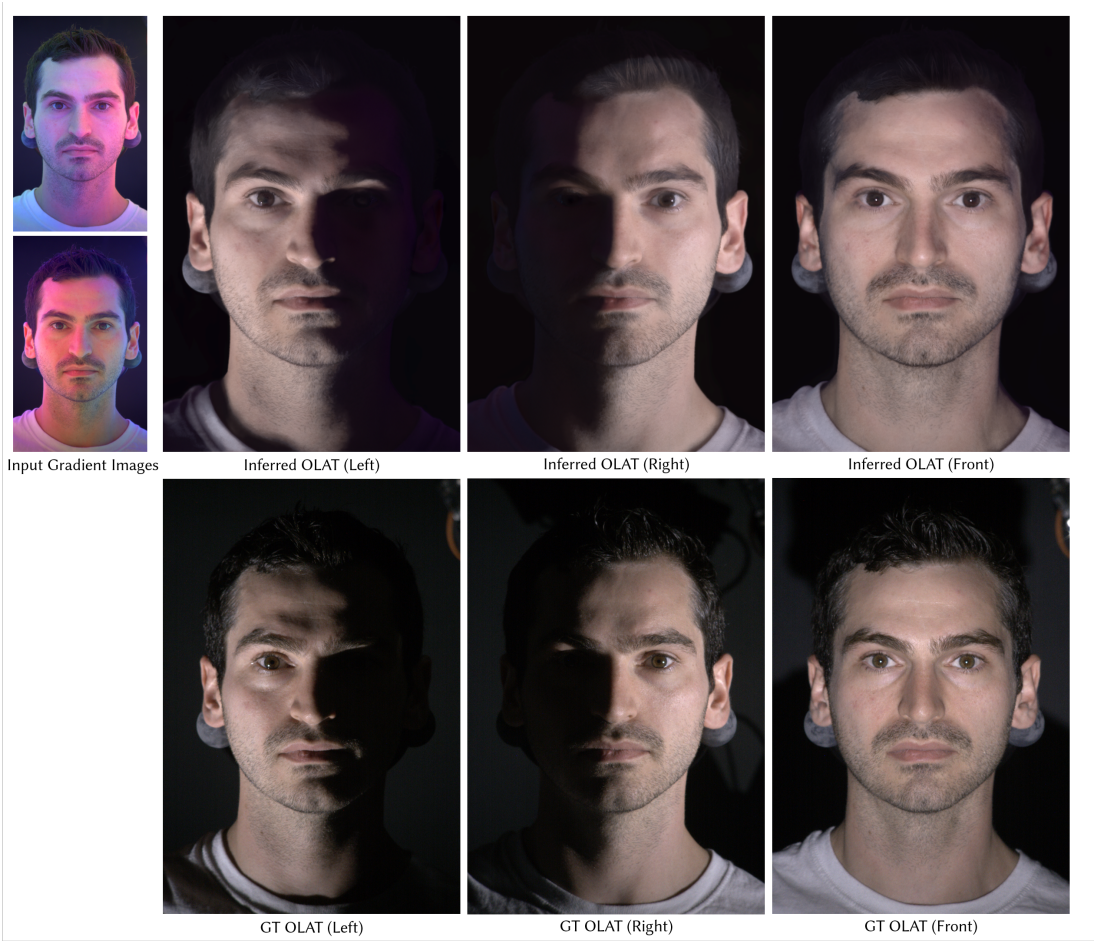


Figure 7.6: **Qualitative results** – examples of gradient input images, inferred OLAT images and ground truth. Notice how high-frequency details such as specularities, shadows and skin texture are correctly extracted from the gradient images.

Viewpoint generalization In Figure 7.8, generalization with respect to viewpoints is demonstrated. As discussed, the rig includes 9 cameras of which only 5 are used for training, leaving 4 unseen viewpoints for testing. As it can be observed, the proposed method does not introduce any specific artifacts with respect to the viewpoint. This demonstrates that the gradient images contain enough information so that the network can infer some notion of geometry.

Comparison with the state-of-the-art In Figure 7.9, the results from the proposed method are compared to the state-of-the-art approaches of Fyffe *et al.* (2009) and Shu *et al.* (2017). The method of Fyffe *et al.* (2009) also takes as input two gradient illumination images, but relies on the cosine lobe reflectance model

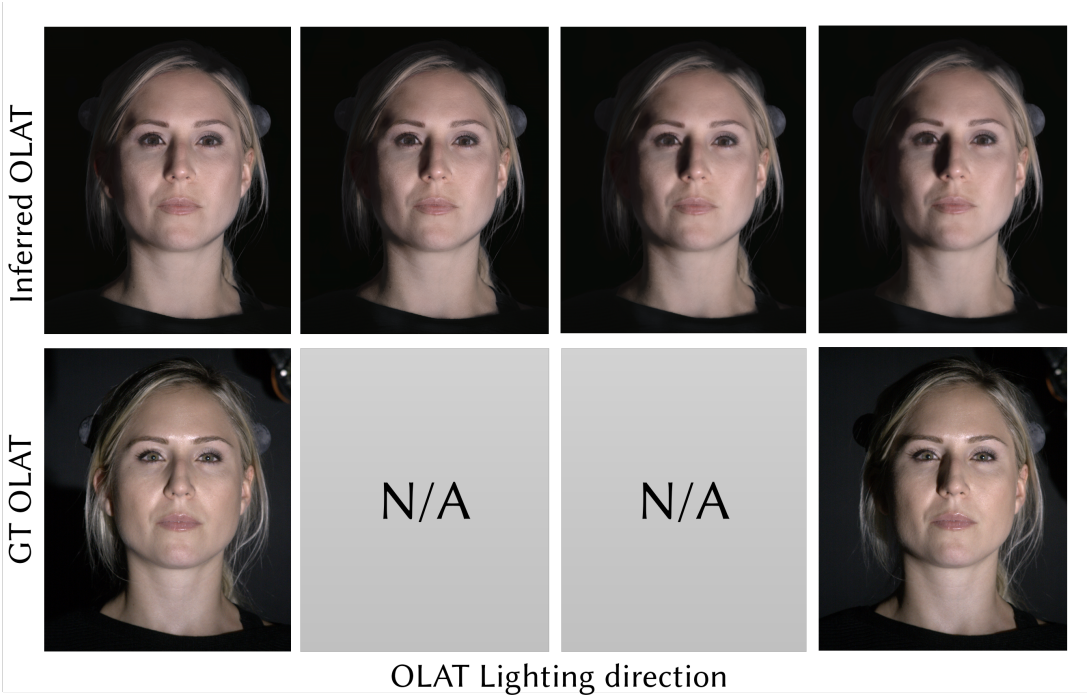


Figure 7.7: **Generalization w.r.t. light direction** – Top row: inferred OLAT images; bottom row: ground truth OLAT images. The centre two columns correspond to two lighting directions interpolated between the lighting directions in the far left and far right columns. Despite, not having ground truth images for these directions, due to the sparsity of lights on the Light Stage, the proposed method can infer OLAT images (top middle images).

to generate images under arbitrary lighting conditions. Note this method has the drawback that it requires an additional colour correction calibration to account for differences in camera colour primaries vs light colour primaries, whereas the proposed method simply learns to relight in whichever colour space the input data is given (note that in the experiments, due to the missing colour calibration step, there is a purple colour cast in the results). Shu *et al.* (2017) uses a light transport approach: the method transfers lighting from a source portrait image to a target portrait image. It is utilized here to transfer the lighting from on OLAT image of the source subject to the fully lit image of a target subject thereby generating a single OLAT image of the source subject. Conducting the transfer for each source OLAT image allows us to generate all the OLAT images for the target subject from a single fully lit image of the target subject. The final relighting with the environment map is based on the generated OLAT images. Notice how the results produced by the other baselines lack details where the proposed method is able to infer even extreme oblique spotlights.

Figure 7.8: **Generalization w.r.t. viewpoint** – The method is able to generalize across views, showing that the input incorporates some form of geometrical information that the network can exploit.

Dynamic Capture In Figure 7.11, it is shown how the proposed method is able to handle *dynamic* subjects performing arbitrary motions and expressions. Note that no ground-truth is available for these sequences as OLAT ground-truth acquisition is feasible only for *static* scenes. Importantly, the trained network is able to generalize to facial expressions which are not present in the training data, which is captured with a neutral expression. Compared to Fyffe *et al.* (2009), the technique produces more natural skin reflectance, a better reproduction of specular highlights, and significantly better shadows.

Figure 7.9: Comparison of OLAT images – OLAT images generated with different methods are compared.

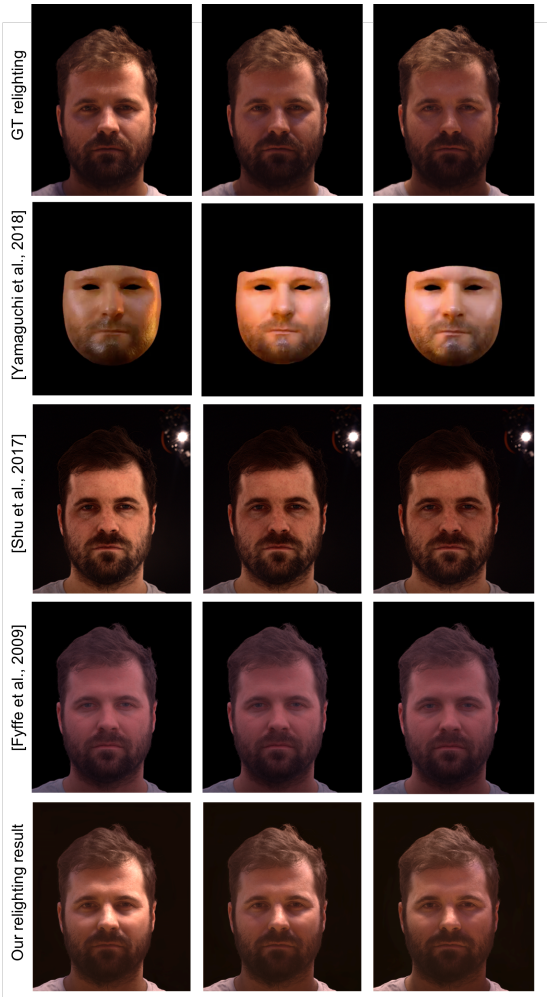


Figure 7.10: Relighting Comparisons – Comparisons with other state of art approaches. The proposed method outperforms the latest work in the field.

7.5.2 Ablation study

In Figure 7.4, the effect of each component of the proposed loss function is evaluated. The proposed loss outperforms a VGG network pre-trained on ImageNet that is not able to pick up specularities, shadows and high-frequency detail. Furthermore, the alignment strategy proposed in Equation (7.2) leads to sharper results.

In Figure 7.12, different input modalities are explored by training a network that takes as input a subset of the OLAT images (with wide and narrow baseline between the input lighting directions) and infers the remaining ones. Note how

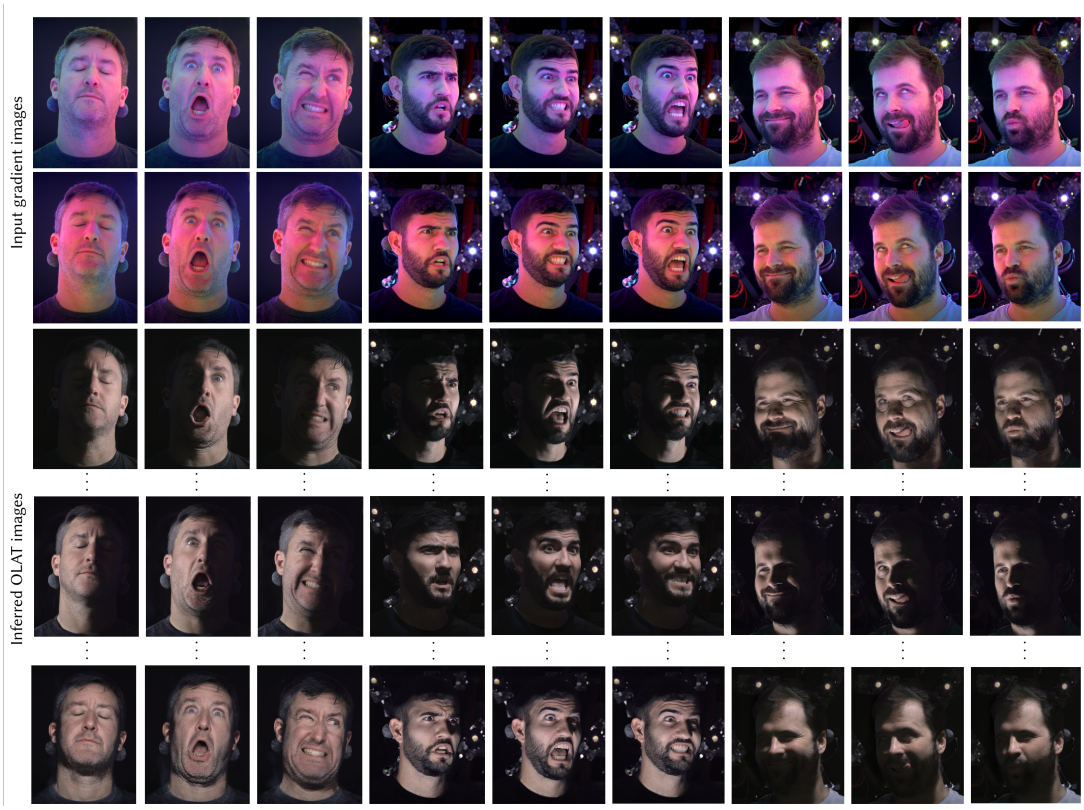


Figure 7.11: **Dynamic Capture** – Top rows: input gradients for a moving subject. Bottom rows: inferred OLAT images.

	Proposed	$\mathbf{L}_{\text{pretrained}}$	$\mathbf{L}_{\text{specific}}$	No Alignment Loss	3-OLAT Input
Photometric Error	**808.64**	917.82	914.81	956.98	1320.51
MS-SSIM	0.222	0.217	0.216	**0.290**	0.216

Table 7.1: Quantitative evaluations on test sequences of subjects. Photometric error is measured via the ℓ_1-norm. Keeping architecture fixed, the proposed loss function is compared with the other baselines. Significantly lower MSE with the ground-truth is obtained while the SSIM score is similar to the other networks. Do note that these statistical measured often do not quantify well the subjective photorealism of the images.

these networks failed to recover high-frequency shadows and texture, proving that the proposed gradient images are a better choice for the relighting task.

In Table 7.1, *quantitative* evaluations are reported. In particular metrics such as photometric error and MS-SSIM are computed by training multiple architectures where one or more losses are selectively used.

Figure 7.12: **Ablation study** – Comparisons with different input modalities. Taking three OLAT as input (top) does not perform as well as the proposed gradient images (bottom).

7.5.3 User study

In order to objectively gauge the quality of the predicted OLATs, two user studies are executed, one with static images, and the other with videos. The first user study consisted of 10 randomly sampled ground-truth OLAT images and 10 images predicted by the proposed network for 140 users to assess. The users are shown each OLAT image and with no additional information, they are asked if they believe that the image is real, i.e., captured using an actual camera, or synthetic. Among the 2800 responses, participants were able to correctly identify the real or synthetic images 79% of the time, indicating that there is room for improvement of the quality of the OLAT images generated by the proposed method. Among the wrong assessments, 50.8% of the real images were wrongly determined to be synthetic and

Figure 7.13: **Relighting with HDRI lighting environments** – Row 1: Ground truth OLAT base relighting, Row 2: cosine lobe relighting Fyffe *et al.* (2009), Row 3: relighting results from the proposed method. Notice how the proposed method outperforms all state-of-the-art methods and comes very close to the ground truth.

49.2% of the synthetic images were wrongly classified as real.

In the second user study with 58 participants, users were shown 6 randomly selected video relighting results, to gauge whether the video appeared real or fake. Of the net 348 responses, more than 66% were marked real. This shows that even though the inferred OLATs might appear synthetic, the relighting results under high-frequency environment maps were mostly considered realistic. The participants were also asked what *cues* they used to decide if an image was real or synthetic. Common responses included issues with eye highlights, teeth texture and general blurriness.

7.5.4 Environment map based relighting

The estimated reflectance field is used to relight an image under a new lighting environment. Given an environment map, OLAT images are used to produce an image under the desired illumination, see Figure 7.2. For each OLAT image an RGB scaling factor is assigned based on the intensity of the environment map as in Debevec *et al.* (2000). The final relit image is generated as a linear combination of the

weighted OLAT images. In Figure 7.13, comparison with Fyffe *et al.* (2009) is shown, which uses exactly the same input but requires a prior in the form of a parametric BRDF representation. In Figure 7.10, comparisons with the relighting capability of the state-of-the-art works of Shu *et al.* (2017) and Yamaguchi *et al.* (2018) are shown. The results from the proposed method show state-of-the-art quality for a method that can perform relighting of dynamic sequences *without* resorting to parametric priors.

7.6 Discussion

While the results are generally realistic and are a significant improvement over existing techniques, analysis shows that the synthesized OLAT images sometimes exhibit small artifacts and occasional over-smoothing. Specular highlights in the estimated OLAT images are often attenuated, and at times missing on particularly noticeable regions of the face, such as eyes and teeth. This may be a result of the low-frequency nature of the colour gradient illumination. This type of illumination captures enough information to infer a simple reflectance model that does not capture high-frequency effects at full extent. Additional (static) capture sessions could be used to better capture the reflectance of skin. In other words, an interesting question for future research is the relationship between lighting pattern configuration and reflectance information that can be captured.

As common in time-multiplexed capture, the slight misalignment between the two input images causes temporal artifacts. It is expected that the continuous improvements in camera hardware will help mitigate these issues. Nonetheless, note that in environment relighting results these artifacts are averaged out by the integration process, resulting in plausible relighting.

7.7 Conclusion

A novel approach for capturing high-quality time-varying 4D reflectance field at 30 FPS has been proposed which does not require high-speed cameras, motion compensation, or parametric priors. This enables a high-quality method to generate relightable dynamic sequences of human actors. The proposed approach provides a simple and effective model and process which can be applied not only to producing high-quality time-varying 4D reflectance fields of faces, but potentially to *any* static or dynamic object.

7.7 Conclusion

It remains of interest to improve the output quality while making the algorithm more efficient. An interesting way to improve performance is to reuse features extracted early on in the network for all the OLAT directions one would predict, for example by using a late fusion technique for the lighting direction instead of the current early fusion. To further improve output quality, novel task-specific perceptual losses and neural generative techniques (GANs) can be explored to further aid in recovering high-frequency details. An additional interesting avenue of research is to explore input representations beyond spherical gradient images which could lead to higher quality outputs. Such representations could be handcrafted, or even learned as part of the neural network training process. The success of the proposed method in recovering detailed reflectance fields suggests the tantalizing possibility of high-quality multi-view geometry and BRDF capture along the lines of Ghosh *et al.* (2011); Ma *et al.* (2007) in dynamic settings.

This concludes the chapters discussing the technical work performed for this thesis. In the next chapter, an analysis of the breadth of this research work and ideas for future research directions are presented.

Chapter 8

Conclusion

This thesis explores various solutions to the inverse rendering problem in live scenarios, with a minimal number of cameras, in unconstrained settings and generalizing across various classes of objects and scenes.

Beginning with the single-light-bounce Lambertian reflectance case, each chapter in this thesis adds a level of complexity by dropping simplifying assumptions about the nature of the scene reflectance and lighting. A range of new representations and algorithms are presented for tackling these under-constrained high-dimensional and non-linear problems. A particular emphasis is laid on an extensive qualitative and quantitative analysis that establishes the efficacy of the proposed methods.

The thesis demonstrates several novel applications in real-time image/video editing and photorealistic mixed reality space such as dynamic-lighting-aware retexturing, global-illumination-aware recolouring, material-editing, material-transfer and high-quality and high-resolution facial relighting.

8.1 Insights

Aside from these contributions, several broader insights arise from the breadth of all the projects discussed in this thesis.

Low-dimensional geometry encoding aids in generalizing inverse rendering algorithms across geometry classes. Geometry, reflectance and illumination are intricately involved in the process of image formation as seen in the rendering equation (Equation (2.1)). Thus, the problem of estimating one of these quantities from 2D data inherently involves solving for the others. But as was noted in Chapter 6 and Chapter 7, generalizing a reflectance or lighting estimation method

across diverse geometry classes is non-trivial. In this thesis work, shading or gradient images, that encode information about local curvature (normals) and local irradiance, were used as a proxy for geometry, in order to simplify the formulation. This was significant in enabling the methods to generalize across geometry classes for two reasons. First, it inherently prevents the method from being over-designed or over-fit to a geometry reconstruction method that may itself be inaccurate or biased towards certain classes of objects. Second, it allows for easy handling of geometry since solving for 3D geometry on the regular 2D grid of the GPU cores or 2D convolutional neural networks has several challenges. Following this observation, finding simpler geometry encodings could be key to the generalization problem in inverse rendering.

Joint solutions to the various light transport components enhance the quality of individual estimates. In Chapter 5 it was shown that estimating the 2nd bounce of light (indirect shading layer) in a scene not only provides additional information about the light transport, it also improves the quality of the reflectance (base colours) and the 1st light bounce (direct shading) estimates as compared to earlier methods in Chapter 3 and Chapter 4. In Chapter 6, designing the discriminative approach to estimate a high-frequency lighting map of the environment reflected onto the object in the form of the 'mirror-image' led to an improved estimate of the specular 'shininess' exponent. In Chapter 7, adding a task-specific loss that was designed to regress the lighting direction from an image patch resulted in significantly improved reflectance field estimates. From these, it is evident that formulations that are more comprehensive in modelling or accounting for the various light transport components lead to an enhancement in the quality of the estimation of the primary components.

Learning light transport components for distinct sub-spaces is beneficial. In the past, a great deal of effort has gone into designing reflectance and illumination models that can simulate natural materials (Matusik *et al.*, 2003) and lighting (Georgoulis *et al.*, 2016, 2018; Lombardi & Nishino, 2016a). The search for these natural sub-spaces – within the space of all possible 4D BRDF functions or 3D lighting functions – still remains the holy grail of inverse rendering. In the era of data-driven learning approaches, novel representations for distinct underlying distributions of reflectance and lighting, such that they are directly amenable to applications such as rendering, are very useful. As was evidenced in Chapter 6, working exclusively with training data of indoor environment maps allowed the networks to learn to interpret the effects of high-frequency light sources in an image to compute reflectance. In Chapter 7, an explicit reflectance basis was learnt based

on directional light sources, and this basis could be used directly to render relighting results. Thus, dealing with specific sub-spaces of reflectance and lighting to develop representations that are directly applicable has been shown to be advantageous.

8.2 Outlook

While this thesis *begins* probing the space of live inverse rendering problems in casual settings, it also suggests several interesting challenges that are forthcoming in the field.

The advances in deep learning methods in the recent past offer several opportunities. There are many classes of problems, such as inverse rendering in indoor settings (Gardner *et al.*, 2017; Garon *et al.*, 2019; Li *et al.*, 2019), outdoor settings (Hold-Geoffroy *et al.*, 2017; Yu & Smith, 2018)), using objects as light probes (Calian *et al.*, 2018; Weber *et al.*, 2018), image based relighting (Philip *et al.*, 2019; Xu *et al.*, 2018), sv-BRDF estimation of material samples (Deschaintre *et al.*, 2018, 2019; Li *et al.*, 2018a,b) that have greatly benefited from learning from data.

Some directions for future work are discussed below.

Learning implicit representations

Traditional low-dimensional models for reflectance and lighting such as the Blinn-Phong model or the spherical harmonics lighting used in this thesis are analytical. While they have been shown to be useful for simplifying the problem formulations, they are not amenable to ray-tracing. The discriminative approaches of Chapter 6 and Chapter 7 were effective in estimating high-frequency reflectance and illumination, but they left something to be desired in explicitly modelling global illumination effects such as inter-reflections and shadows. Recently, several implicit representations for geometry have been proposed (Lombardi *et al.*, 2019; Sitzmann *et al.*, 2019a,b) that learn differentiable ray-tracers. These latent space representations allow for geometry estimation from single or multi-view 2D images and also enable novel-view synthesis and provide correspondence across views. Since the representation is itself differentiable and can be probed by a learnt ray-tracer, this makes it viable for estimating contributions of light transport effects such as specularities and global illumination in real images by fully inverting the rendering process.

Learning programmable shaders

Complex light transport effects such as global illumination, translucency, specular microgeometry and sub-surface scattering are computationally expensive to generate. Learning shaders that can directly render such effects in image space can be beneficial to achieving photorealism. While there has been some work in this direction recently (Alexandr Kuznetsov, 2019), creating a general class of learnt tunable shaders that can perform such tasks is an unaddressed research problem. By learning to generate such effects, differentiable models and representation can be created that may also be used to estimate complex light transport components from real world images. A layered approach to simulate these effects can be followed, similar to intrinsic image layers described in Chapter 5. Recent advances in reinforcement-learning based methods that deconstruct an image into painting strokes (Huang *et al.*, 2019; Zheng *et al.*, 2019) can be extended to paint semantically meaningful layers in image space.

Using additional consumer hardware

Live inverse rendering is a hard problem primarily because it is heavily under-constrained. This has been the general theme of this thesis. In Chapter 4, a depth sensor was additionally used to solve the inherent ambiguity of texture-copy and in Chapter 6 to estimate environment maps. In Chapter 7, a Light Stage is used to generate special lighting conditions to estimate reflectance fields in real-time for dynamic facial performances. Similarly, additional consumer hardware that is easily available on mobile devices can be cleverly used to obtain better constraints. Recently, the camera flash has been effectively used to improve reflectance estimation for general object classes (Li *et al.*, 2018b) or separate contribution of different illuminants to an image (Hui *et al.*, 2019). A 360° camera has been used as a light probe to improve portrait relighting (E *et al.*, 2019). Along similar lines, readily available hardware such as mobile phone screens can be used as projectors, or polarization filters on cameras used to separate specular effects. These additional signals can provide better constraints to significantly improve the quality of reflectance and lighting estimation using consumer devices.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Man, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Vigas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y. & Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org. 113

Aksoy, Y., Aydn, T.O., Pollefeys, M. & Smoli, A. (2016). Interactive high-quality green-screen keying via color unmixing. *ACM Transactions on Graphics*, **35**, 152:1–12. 72, 74, 83

Aksoy, Y., Aydn, T.O., Smoli, A. & Pollefeys, M. (2017). Unmixing-based soft color segmentation for image manipulation. *ACM Transactions on Graphics*, **36**, 19:1–19. 74

Alexandr Kuznetsov, Z.X.L.Q.Y.B.W.N.K.K.S.M.R.R., Milo Haan (2019). Learning generative models for rendering specular microgeometry. **38**. 148

Anderson, R., Gallup, D., Barron, J.T., Kontkanen, J., Snavely, N., Hernández, C., Agarwal, S. & Seitz, S.M. (2016). Jump: Virtual reality video. *SIGGRAPH Asia*. 130

Azinovi, D., Li, T.M., Kaplanyan, A. & Niener, M. (2019). Inverse path tracing for joint material and lighting estimation. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR)*. 11

REFERENCES

Bach, F., Jenatton, R., Mairal, J. & Obozinski, G. (2012). Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, **4**, 1–106. 82

Barron, J.T. & Malik, J. (2013). Intrinsic scene properties from a single RGB-D image. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR)*, 17–24. 26

Barron, J.T. & Malik, J. (2015a). Shape, illumination, and reflectance from shading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **37**, 1670–1687. 22, 25, 54

Barron, J.T. & Malik, J. (2015b). Shape, illumination, and reflectance from shading. *IEEE Trans. Pattern Anal. Mach. Intell.*, **37**. 125

Barrow, H.G. & Tenenbaum, J.M. (1978a). Recovering intrinsic scene characteristics from images. Tech. Rep. 157, AI Center, SRI International. 16

Barrow, H.G. & Tenenbaum, J.M. (1978b). Recovering intrinsic scene characteristics from images. In A. Hanson & E. Riseman, eds., *Computer Vision Systems*, 3–26. 52

Bell, M. & Freeman, W.T. (2001). Learning local evidence for shading and reflection. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR)*. 24, 25

Bell, S., Bala, K. & Snavely, N. (2014). Intrinsic images in the wild. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, **33**, 159:1–12. xvi, 22, 25, 42, 54, 90, 91, 99

Bérard, P., Bradley, D., Gross, M. & Beeler, T. (2016). Lightweight eye capture using a parametric model. *ACM Trans. Graph. (Proc. SIGGRAPH)*, **35**. 126

Bermano, A., Beeler, T., Kozlov, Y., Bradley, D., Bickel, B. & Gross, M. (2015). Detailed spatio-temporal reconstruction of eyelids. *ACM Trans. Graph.*. 126

Bi, S., Han, X. & Yu, Y. (2015). An l_1 image transform for edge-preserving smoothing and scene-level intrinsic decomposition. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, **34**, 78:1–12. 24, 54

Blanz, V. & Vetter, T. (1999). A morphable model for the synthesis of 3d faces. In *Proc. of the Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '99. 126

Blinn, J.F. (1977). Models of light reflection for computer synthesized pictures. *Computer Graphics (Proceedings of SIGGRAPH)*, **11**, 192–198. 13, 108

Bonneel, N., Sunkavalli, K., Tompkin, J., Sun, D., Paris, S. & Pfister, H. (2014). Interactive intrinsic video editing. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)*, **33**, 197:1–10. 22, 23, 24, 25, 26, 29, 30, 34, 39, 40, 41, 42, 49, 52, 54, 59, 65, 66, 67, 77, 86, 87, 92, 93, 94, 95, 96, 97

Bonneel, N., Tompkin, J., Sunkavalli, K., Sun, D., Paris, S. & Pfister, H. (2015). Blind video temporal consistency. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)*, **34**, 196:1–9. 26

Bousseau, A., Paris, S. & Durand, F. (2009). User-assisted intrinsic images. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)*, **28**, 130:1–10. 25, 54, 59, 72

Calian, D.A., Lalonde, J.F., Gotardo, P., Simon, T., Matthews, I. & Mitchell, K. (2018). From faces to outdoor light probes. *Computer Graphics Forum*, **37**, 51–61. 147

Carroll, R., Ramamoorthi, R. & Agrawala, M. (2011). Illumination decomposition for material recoloring with consistent interreflections. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, **30**, 43:1–10. 72, 76, 77, 79, 82, 83, 87, 94, 95, 99, 106

Chang, J., Cabezas, R. & Fisher III, J.W. (2014). Bayesian nonparametric intrinsic image decomposition. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 24

Chen, J., Bautembach, D. & Izadi, S. (2013). Scalable real-time volumetric surface reconstruction. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, **32**, 113. 55

Chen, Q. & Koltun, V. (2013). A simple model for intrinsic image decomposition with depth cues. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 241–248. 22, 25, 26, 54

Chen, Q. & Koltun, V. (2017). Photographic image synthesis with cascaded refinement networks. *ICCV*. 131

Chollet, F. *et al.* (2015). Keras. https://github.com/fchollet/keras. 113

Curless, B. & Levoy, M. (1996). A volumetric method for building complex models from range images. In *Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*. 55, 56

Dai, A., Niener, M., Zollhfer, M., Izadi, S. & Theobalt, C. (2017). BundleFusion: Real-time globally consistent 3D reconstruction using on-the-fly surface reintegration. *ACM Transactions on Graphics*, **36**, 24:1–18. 105

Debevec, P. (1998). Rendering synthetic objects into real scenes: bridging traditional and image-based graphics with global illumination and high dynamic range photography. In *Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, 189–198. 14, 105, 117

Debevec, P. (2012). The Light Stages and Their Applications to Photoreal Digital Actors. In *SIGGRAPH Asia*, Singapore. 124, 129

Debevec, P., Hawkins, T., Tchou, C., Duiker, H.P., Sarokin, W. & Sagar, M. (2000). Acquiring the reflectance field of a human face. In *Proceedings of SIGGRAPH 2000*, SIGGRAPH '00. 124, 126, 142

Deschaintre, V., Aittala, M., Durand, F., Drettakis, G. & Bousseau, A. (2018). Single-image svbrdf capture with a rendering-aware deep network. *ACM Transactions on Graphics (SIGGRAPH Conference Proceedings)*, **37**, 15. 147

Deschaintre, V., Aittala, M., Durand, F., Drettakis, G. & Bousseau, A. (2019). Flexible svbrdf capture with a multi-image deep network. *Computer Graphics Forum (Proceedings of the Eurographics Symposium on Rendering)*, **38**. 147

Di Renzo, F., Calabrese, C. & Pellacini, F. (2014). AppIm: Linear spaces for image-based appearance editing. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)*, **33**, 194:1–9. 106

Ding, S., Sheng, B., Hou, X., Xie, Z. & Ma, L. (2017). Intrinsic image decomposition using multi-scale measurements and sparsity. *Computer Graphics Forum*, **36**, 251–261. 54

Dong, B., Dong, Y., Tong, X. & Peers, P. (2015). Measurement-based editing of diffuse albedo with consistent interreflections. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, **34**, 112:1–11. 72, 74, 106

Dong, Y., Chen, G., Peers, P., Zhang, J. & Tong, X. (2014). Appearance-from-motion: Recovering spatially varying surface reflectance under unknown lighting. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)*, **33**, 193:1–12. 11, 106

Duchêne, S., Riant, C., Chaurasia, G., Moreno, J.L., Laffont, P.Y., Popov, S., Bousseau, A. & Drettakis, G. (2015). Multiview intrinsic images of outdoors scenes with an application to relighting. *ACM Transactions on Graphics*, **34**, 164:1–16. 22, 25

E, J.L., Fried, O. & Agrawala, M. (2019). Optimizing portrait lighting at capture-time using a 360 camera as a light probe. In *To appear, UIST'19*, ACM. 148

Einarsson, P., Chabert, C.F., Jones, A., Ma, W.C., Lamond, B., Hawkins, T., Bolas, M., Sylwan, S. & Debevec, P. (2006). Relighting human locomotion with flowed reflectance fields. In *Proceedings of the 17th Eurographics Conference on Rendering Techniques*, EGSR '06. 124, 126

Eslami, S.M.A., Jimenez Rezende, D., Besse, F., Viola, F., Morcos, A.S., Garnelo, M., Ruderman, A., Rusu, A.A., Danihelka, I., Gregor, K., Reichert, D.P., Buesing, L., Weber, T., Vinyals, O., Rosenbaum, D., Rabinowitz, N., King, H., Hillier, C., Botvinick, M., Wierstra, D., Kavukcuoglu, K. & Hassabis, D. (2018). Neural scene representation and rendering. *Science*, **360**. 131

Favreau, J.D., Lafarge, F. & Bousseau, A. (2017). Photo2ClipArt: Image abstraction and vectorization using layered linear gradients. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)*, **36**, 180:1–11. 74

Fuhrmann, S. & Goesele, M. (2014). Floating scale surface reconstruction. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, **33**, 46:1–11. 55

Fyffe, G. & Debevec, P. (2015). Single-shot reflectance measurement from polarized color gradient illumination. In *ICCP*. 129

Fyffe, G., Wilson, C.A. & Debevec, P. (2009). Cosine lobe based relighting from gradient illumination photographs. In *SIGGRAPH '09: Posters*, SIGGRAPH '09. 126, 127, 128, 129, 136, 138, 142, 143

Garces, E., Munoz, A., Lopez-Moreno, J. & Gutierrez, D. (2012). Intrinsic images by clustering. *Computer Graphics Forum (Proceedings of Eurographics Symposium on Rendering)*, **31**, 1415–1424. 24, 54, 77

Gardner, M.A., Sunkavalli, K., Yumer, E., Shen, X., Gambaretto, E., Gagn, C. & Lalonde, J.F. (2017). Learning to predict indoor illumination from a single image. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)*, **36**, 176:1–14. 106, 147

Garon, M., Sunkavalli, K., Hadap, S., Carr, N. & Lalonde, J.F. (2019). Fast spatially-varying indoor lighting estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 147

Garrido, P., Valgaert, L., Wu, C. & Theobalt, C. (2013). Reconstructing detailed dynamic face geometry from monocular video. *ACM Trans. Graph. (Proc. SIGGRAPH Asia)*, **32**. 126

Garrido, P., Zollhoefer, M., Casas, D., Valgaerts, L., Varanasi, K., Perez, P. & Theobalt, C. (2016). Reconstruction of personalized 3d face rigs from monocular video. 126

Gehler, P.V., Rother, C., Kiefel, M., Zhang, L. & Schlkopf, B. (2011). Recovering intrinsic images with a global sparsity prior on reflectance. In *Advances in Neural Information Processing Systems*. 22, 24, 53, 54

Georgoulis, S., Rematas, K., Ritschel, T., Fritz, M., Gool, L.V. & Tuytelaars, T. (2016). DeLight-net: Decomposing reflectance maps into specular materials and natural illumination, arXiv:1603.08240. 146

Georgoulis, S., Rematas, K., Ritschel, T., Fritz, M., Tuytelaars, T. & Gool, L.V. (2017a). What is around the camera? In *Proceedings of the International Conference on Computer Vision (ICCV)*. 106

Georgoulis, S., Rematas, K., Ritschel, T., Gavves, E., Fritz, M., Gool, L.V. & Tuytelaars, T. (2017b). Reflectance and natural illumination from single-material specular objects using deep learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **preprints**. 11, 104, 105, 106, 107, 111, 116, 118, 119

Georgoulis, S., Rematas, K., Ritschel, T., Gavves, E., Fritz, M., Van Gool, L. & Tuytelaars, T. (2018). Reflectance and natural illumination from single-material specular objects using deep learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **40**, 1932–1947. 146

Ghosh, A., Fyffe, G., Tunwattanapong, B., Busch, J., Yu, X. & Debevec, P. (2011). Multiview face capture using polarized spherical gradient illumination. *ACM Trans. Graph.*. 144

Gong, W., Zhang, X., Gonzlez, J., Sobral, A., Bouwmans, T., Tu, C. & ZAHZAH, E.h. (2016). Human pose estimation from monocular images: A comprehensive survey. *Sensors*, **16**, 1966. 3

Gotardo, P., Riviere, J., Bradley, D., Ghosh, A. & Beeler, T. (2018). Practical dynamic facial appearance modeling and acquisition. In *SIGGRAPH Asia*. 126

Grosse, R., Johnson, M.K., Adelson, E.H. & Freeman, W.T. (2009). Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2335–2342. 24, 40, 87, 88

Gruber, L., Richter-Trummer, T. & Schmalstieg, D. (2012). Real-time photometric registration from arbitrary geometry. In *Proceedings of the International Symposium on Mixed and Augmented Reality (ISMAR)*. 54

Gruber, L., Ventura, J. & Schmalstieg, D. (2015). Image-space illumination for augmented reality in dynamic environments. In *Proceedings of IEEE Virtual Reality (VR)*, 127–134. 55

Guo, K., Xu, F., Yu, T., Liu, X., Dai, Q. & Liu, Y. (2017). Real-time geometry, albedo and motion reconstruction using a single RGBD camera. *ACM Transactions on Graphics*, **36**, 32:1–13. 11, 105

Hachama, M., Ghanem, B. & Wonka, P. (2015). Intrinsic scene decomposition from RGB-D images. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 810–818. 26, 54

Han, X., Laga, H. & Bennamoun, M. (2019). Image-based 3d object reconstruction: State-of-the-art and trends in the deep learning era. *CoRR*, **abs/1906.06543**. 11

Hartley, R. & Zisserman, A. (2003). *Multiple View Geometry in Computer Vision*. Cambridge University Press, New York, NY, USA, 2nd edn. 11

Hauagge, D., Wehrwein, S., Bala, K. & Snavely, N. (2013). Photometric ambient occlusion. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2515–2522. 25

Hawkins, T., Wenger, A., Tchou, C., Gardner, A., Göransson, F. & Debevec, P.E. (2004). Animatable facial reflectance fields. *Rendering Techniques*. 126

Hold-Geoffroy, Y., Sunkavalli, K., Hadap, S., Gambaretto, E. & Lalonde, J.F. (2017). Deep outdoor illumination estimation. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR)*. 106, 147

Holland, P.W. & Welsch, R.E. (1977). Robust regression using iteratively reweighted least-squares. *Communications in Statistics Theory and Methods*, **6**, 813–827. 34, 84

Horn, B.K.P. (1974). Determining lightness from an image. *Computer Graphics and Image Processing*, **3**, 277–299. 22

Hu, L., Ma, C., Luo, L. & Li, H. (2015). Single-view hair modeling using a hairstyle database. *ACM Trans. on Graphics (SIGGRAPH)*. 126

Huang, Z., Heng, W. & Zhou, S. (2019). Learning to paint with model-based deep reinforcement learning. *CoRR*, **abs/1903.04411**. 148

Hui, Z. & Sankaranarayanan, A.C. (2017). Shape and spatially-varying reflectance estimation from virtual exemplars. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**, 2060–2073. 105

Hui, Z., Chakrabarti, A., Sunkavalli, K. & Sankaranarayanan, A.C. (2019). Learning to separate multiple illuminants in a single image. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 148

Ichim, A.E., Bouaziz, S. & Pauly, M. (2015). Dynamic 3d avatar creation from hand-held video input. *ACM Trans. Graph.*, **34**. 126

Immel, D.S., Cohen, M.F. & Greenberg, D.P. (1986). A radiosity method for non-diffuse environments. *SIGGRAPH Comput. Graph.*, **20**, 133–142. 9

Innamorati, C., Ritschel, T., Weyrich, T. & Mitra, N.J. (2017). Decomposing single images for layered photo retouching. *Computer Graphics Forum (Proceedings of Eurographics Symposium on Rendering)*, **36**, 15–25. 74

Innmann, M., Zollhfer, M., Niener, M., Theobalt, C. & Stamminger, M. (2016). VolumeDeform: Real-time volumetric non-rigid reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 105

Isola, P., Zhu, J.Y., Zhou, T. & Efros, A.A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR)*, 5967–5976. 107, 131

Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A. & Fitzgibbon, A. (2011). KinectFusion: real-time 3D reconstruction and interaction using a moving depth camera. In *Proceedings of the Symposium on User Interface Software and Technology (UIST)*, 559–568. 55, 105

Jiang, X., Schofield, A.J. & Wyatt, J.L. (2010). Correlation-based intrinsic image extraction from a single image. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 24

Joshi, N., Zitnick, C., Szeliski, R. & Kriegman, D. (2009). Image deblurring and denoising using color priors. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR)*, 1550–1557. 34

Kajiya, J.T. (1986). The rendering equation. *Computer Graphics (Proceedings of SIGGRAPH)*, **20**, 143–150. 2, 9

Kanamori, Y. & Endo, Y. (2018). Relighting humans: Occlusion-aware inverse rendering for full-body human images. In *SIGGRAPH Asia*, ACM. 127

Kerl, C., Souiai, M., Sturm, J. & Cremers, D. (2014). Towards illumination-invariant 3D reconstruction using ToF RGB-D cameras. In *Proceedings of International Conference on 3D Vision (3DV)*. 55

Khan, E.A., Reinhard, E., Fleming, R.W. & Bülthoff, H.H. (2006). Image-based material editing. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, **25**, 654–663. 106

Kim, K., Gu, J., Tyree, S., Molchanov, P., Niener, M. & Kautz, J. (2017). A lightweight approach for on-the-fly reflectance estimation. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 20–28. 104, 105, 106

Kingma, D.P. & Ba, J. (2015). Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*. 113, 132

Klein, G. & Murray, D. (2007). Parallel tracking and mapping for small AR workspaces. In *Proceedings of the International Symposium on Mixed and Augmented Reality (ISMAR)*. 47

Knecht, M., Tanzmeister, G., Traxler, C. & Wimmer, M. (2012). Interactive BRDF estimation for mixed-reality applications. *Journal of WSCG*, **20**, 47–56. 106

Kong, N. & Black, M.J. (2015). Intrinsic depth: Improving depth transfer with intrinsic images. In *Proceedings of the International Conference on Computer Vision (ICCV)*. 54

Kong, N., Gehler, P.V. & Black, M.J. (2014). Intrinsic video. In D. Fleet, T. Pajdla, B. Schiele & T. Tuytelaars, eds., *Proceedings of the European Conference on Computer Vision (ECCV)*, vol. 8690 of *Lecture Notes in Computer Science*, 360–375. 22, 25, 54

Kuo, W., Häne, C., Yuh, E., Mukherjee, P. & Malik, J. (2018). Patchfcn for intracranial hemorrhage detection. *arXiv preprint arXiv:1806.03265*. 132

Laffont, P.Y. & Bazin, J.C. (2015). Intrinsic decomposition of image sequences from local temporal variations. In *Proceedings of the International Conference on Computer Vision (ICCV)*. 25, 54

Laffont, P.Y., Bousseau, A., Paris, S., Durand, F. & Drettakis, G. (2012). Coherent intrinsic images from photo collections. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)*, **31**, 202:1–11. 25, 74

Laffont, P.Y., Bousseau, A. & Drettakis, G. (2013). Rich intrinsic image decomposition of outdoor scenes from multiple views. *IEEE Transactions on Visualization and Computer Graphics*, **19**, 210–224. 22, 25

Lalonde, J.F. & Matthews, I. (2014). Lighting estimation in outdoor image collections. In *Proceedings of International Conference on 3D Vision (3DV)*, 131–138. 106

Lalonde, J.F., Efros, A.A. & Narasimhan, S.G. (2012). Estimating the natural illumination conditions from a single outdoor image. *International Journal of Computer Vision*, **98**, 123–145. 106

Land, E.H. & McCann, J.J. (1971). Lightness and retinex theory. *Journal of the Optical Society of America*, **61**, 1–11. 24, 29, 53

Lee, K., Zhao, Q., Tong, X., Gong, M., Izadi, S., Lee, S., Tan, P. & Lin, S. (2012). Estimation of intrinsic image sequences from image+depth video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, vol. 7577 of *Lecture Notes in Computer Science*, 327–340. 22, 26, 54

Levin, A. & Weiss, Y. (2007). User assisted separation of reflections from a single image using a sparsity prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **29**, 1647–1654. 34

Levin, A., Fergus, R., Durand, F. & Freeman, W.T. (2007). Image and depth from a conventional camera with a coded aperture. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, **26**, 70. 34

Li, G., Wu, C., Stoll, C., Liu, Y., Varanasi, K., Dai, Q. & Theobalt, C. (2013). Capturing Relightable Human Performances under General Uncontrolled Illumination. *Computer Graphics Forum (Proc. EUROGRAPHICS 2013)*. 126

Li, X., Dong, Y., Peers, P. & Tong, X. (2017). Modeling surface appearance from a single photograph using self-augmented convolutional neural networks. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, **36**, 45:1–11. 11, 106

Li, Y. & Brown, M.S. (2014). Single image layer separation using relative smoothness. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2752–2759. 22

Li, Z., Sunkavalli, K. & Chandraker, M. (2018a). Materials for masses: SVBRDF acquisition with a single mobile phone image. In *ECCV*, Lecture Notes in Computer Science, Springer. 126, 147

Li, Z., Xu, Z., Ramamoorthi, R., Sunkavalli, K. & Chandraker, M. (2018b). Learning to reconstruct shape and spatially-varying reflectance from a single image. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)*, **37**, 269:1–11. 11, 126, 147, 148

Li, Z., Shafiei, M., Ramamoorthi, R., Sunkavalli, K. & Chandraker, M. (2019). Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and SVBRDF from a single image. *CoRR*, **abs/1905.02722**. 147

Liao, Z., Karsch, K. & Forsyth, D. (2015). An approximate shading model for object relighting. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR)*, 5307–5314. 106

Liu, G., Ceylan, D., Yumer, E., Yang, J. & Lien, J.M. (2017). Material editing using a physically based rendering network. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2280–2288. 11, 104, 105, 106, 107, 113, 116, 118, 119

Lombardi, S. & Nishino, K. (2016a). Radiometric scene decomposition: Scene reflectance, illumination, and geometry from RGB-D images. In *Proceedings of International Conference on 3D Vision (3DV)*, 305–313. 146

Lombardi, S. & Nishino, K. (2016b). Reflectance and illumination recovery in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **38**, 129–141. 11, 106, 115, 116, 118, 119

Lombardi, S., Saragih, J., Simon, T. & Sheikh, Y. (2018). Deep appearance models for face rendering. *ACM Trans. Graph.*, **37**. 127

Lombardi, S., Simon, T., Saragih, J., Schwartz, G., Lehrmann, A. & Sheikh, Y. (2019). Neural volumes: Learning dynamic renderable volumes from images. *ACM Trans. Graph.*, **38**, 65:1–65:14. 147

Long, J., Shelhamer, E. & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 131

Lopez-Moreno, J., Garces, E., Hadap, S., Reinhard, E. & Gutierrez, D. (2013). Multiple light source estimation in a single image. *Computer Graphics Forum*, **32**, 170–182. 106

Ma, W.C., Hawkins, T., Peers, P., Chabert, C.F., Weiss, M. & Debevec, P. (2007). Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination. In *Proceedings of the Eurographics Conference on Rendering Techniques*, EGSR'07. 144

Mandl, D., Yi, K.M., Mohr, P., Roth, P., Fua, P., Lepetit, V., Schmalstieg, D. & Kalkofen, D. (2017). Learning lightprobes for mixed reality illumination. In *Proceedings of the International Symposium on Mixed and Augmented Reality (ISMAR)*. 103, 106

Marchand, E., Uchiyama, H. & Spindler, F. (2016). Pose estimation for augmented reality: A hands-on survey. *IEEE Transactions on Visualization and Computer Graphics*, **22**, 2633–2651. 3

Marschner, S.R. & Greenberg, D.P. (1997). Inverse lighting for photography. In *Proceedings of the IS&T Color Imaging Conference*, 262–265. 11, 106

Martin-Brualla, R., Pandey, R., Yang, S., Pidlypenskyi, P., Taylor, J., Valentin, J., Khamis, S., Davidson, P., Tkach, A., Lincoln, P., Kowdle, A., Rhemann, C., Goldman, D.B., Keskin, C., Seitz, S., Izadi, S. & Fanello, S. (2018). Lookingood: Enhancing performance capture with real-time neuralre-rendering. In *SIGGRAPH Asia*. 131, 133

Matsushita, Y., Lin, S., Kang, S. & Shum, H.Y. (2004). Estimating intrinsic images from image sequences with biased illumination. In *Proceedings of the European Conference on Computer Vision (ECCV)*, vol. 3022 of *Lecture Notes in Computer Science*, 274–286. 25, 54

Matusik, W., Pfister, H., Brand, M. & McMillan, L. (2003). A data-driven reflectance model. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, **22**, 759–769. 146

Meka, A., Zollhfer, M., Richardt, C. & Theobalt, C. (2016). Live intrinsic video. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, **35**, 109:1–14. 6, 21, 65, 92, 95, 96, 97, 98

Meka, A., Fox, G., Zollhöfer, M., Richardt, C. & Theobalt, C. (2017a). Live user-guided intrinsic video for static scene. *IEEE Transactions on Visualization and Computer Graphics*, **23**. 51

Meka, A., Fox, G., Zollhfer, M., Richardt, C. & Theobalt, C. (2017b). Live user-guided intrinsic video for static scenes. *IEEE Transactions on Visualization and Computer Graphics*, **23**, 2447–2454. 6, 106

Meka, A., Maximov, M., Zollhoefer, M., Chatterjee, A., Seidel, H.P., Richardt, C. & Theobalt, C. (2018). Lime: Live intrinsic material estimation. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*. 6, 103, 127

Meka, A., Haene, C., Pandey, R., Zollhoefer, M., Fanello, S., Fyffe, G., Kowdle, A., Yu, X., Busch, J., Dourgarian, J., Denny, P., Bouaziz, S., Lincoln, P., Whalen, M., Harvey, G., Taylor, J., Izadi, S., Tagliasacchi, A., Debevec, P., Theobalt, C., Valentin, J. & Rhemann, C. (2019a). Deep reflectance fields - high-quality facial reflectance field inference from color gradient illumination. vol. 38. 7, 123

Meka, A., Shafiei, M., Zollhoefer, M., Richardt, C. & Theobalt, C. (2019b). Live illumination decomposition of videos. 6, 71

Moons, T., Van Gool, L. & Vergauwen, M. (2010). *3D Reconstruction from Multiple Images: Part 1: Principles*. now. 11

Nalbach, O., Arabadzhiyska, E., Mehta, D., Seidel, H.P. & Ritschel, T. (2017). Deep shading: Convolutional neural networks for screen-space shading. **36**. 127

Nam, G., Lee, J.H., Gutierrez, D. & Kim, M.H. (2018). Practical SVBRDF acquisition of 3D objects with unstructured flash photography. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)*, **37**, 267:1–12. 11, 126

Nayar, S.K., Krishnan, G., Grossberg, M.D. & Raskar, R. (2006). Fast separation of direct and global components of a scene using high frequency illumination. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, **25**, 935–944. 72, 73

Newcombe, R.A., Davison, A.J., Izadi, S., Kohli, P., Hilliges, O., Shotton, J., Molyneaux, D., Hodges, S., Kim, D. & Fitzgibbon, A. (2011). KinectFusion: Real-time dense surface mapping and tracking. In *Proceedings of the International Symposium on Mixed and Augmented Reality (ISMAR)*, 127–136. 55

Ngan, A., Durand, F. & Matusik, W. (2004). Experimental validation of analytical brdf models. In *ACM SIGGRAPH 2004 Sketches*, SIGGRAPH '04, 90–, ACM, New York, NY, USA. 13

Niener, M., Zollhfer, M., Izadi, S. & Stamminger, M. (2013). Real-time 3D reconstruction at scale using voxel hashing. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)*, **32**, 169:1–11. 55, 56, 64, 66, 105, 120

O'Toole, M., Mather, J. & Kutulakos, K.N. (2016). 3D shape and indirect appearance by structured light transport. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **38**, 1298–1312. 73

Patow, G. & Pueyo, X. (2003). A survey of inverse rendering problems. *Computer Graphics Forum*, **22**, 663–687. 11

Peers, P., Tamura, N., Matusik, W. & Debevec, P. (2007). Post-production facial performance relighting using reflectance transfer. In *ACM SIGGRAPH 2007 Papers*, SIGGRAPH '07, ACM. 126

Philip, J., Gharbi, M., Zhou, T., Efros, A. & Drettakis, G. (2019). Multi-view relighting using a geometry-aware network. *ACM Transactions on Graphics (SIGGRAPH Conference Proceedings)*, **38**. 147

Phong, B.T. (1975). Illumination for computer generated pictures. *Communications of the ACM*, **18**, 311–317. 12

Ramamoorthi, R. & Hanrahan, P. (2001a). An efficient representation for irradiance environment maps. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '01, 497–500, ACM, New York, NY, USA. 14, 16

Ramamoorthi, R. & Hanrahan, P. (2001b). On the relationship between radiance and irradiance: determining the illumination from images of a convex lambertian object. *J. Opt. Soc. Am. A*, **18**, 2448–2459. 16

Ramamoorthi, R. & Hanrahan, P. (2001c). A signal-processing framework for inverse rendering. In *Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, 117–128. 11, 105

Rematas, K., Ritschel, T., Fritz, M., Gavves, E. & Tuytelaars, T. (2016). Deep reflectance maps. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR)*, 4508–4516. xvii, 106, 114, 116

Ren, P., Dong, Y., Lin, S., Tong, X. & Guo, B. (2015). Image based relighting using neural networks. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, **34**, 111:1–12. 74, 127

Richardt, C., Stoll, C., Dodgson, N.A., Seidel, H.P. & Theobalt, C. (2012). Coherent spatiotemporal filtering, upsampling and rendering of RGBZ videos. *Computer Graphics Forum (Proceedings of Eurographics)*, **31**, 247–256. 55

Richardt, C., Lopez-Moreno, J., Bousseau, A., Agrawala, M. & Drettakis, G. (2014). Vectorising bitmaps into semi-transparent gradient layers. *Computer Graphics Forum (Proceedings of Eurographics Symposium on Rendering)*, **33**, 11–19. 74

Richter-Trummer, T., Kalkofen, D., Park, J. & Schmalstieg, D. (2016). Instant mixed reality lighting from casual scanning. In *Proceedings of the International Symposium on Mixed and Augmented Reality (ISMAR)*, 27–36. 106

Rohmer, K., Jendersie, J. & Grosch, T. (2017). Natural environment illumination: Coherent interactive augmented reality for mobile and non-mobile devices. *IEEE Transactions on Visualization and Computer Graphics*, **23**, 2474–2484. 105

Ronneberger, O., Fischer, P. & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W.M. Wells & A.F. Frangi, eds., *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 234–241. 107, 111, 131

Roth, H. & Vona, M. (2012). Moving volume KinectFusion. In *Proceedings of the British Machine Vision Conference (BMVC)*. 55

Rusinkiewicz, S. & Levoy, M. (2001). Efficient variants of the ICP algorithm. In *Proceedings of the International Workshop on 3D Digital Imaging and Modeling (3DIM)*, 145–152. 55

Saito, S., Wei, L., Hu, L., Nagano, K. & Li, H. (2017). Photorealistic facial texture inference using deep neural networks. In *CVPR*, 2326–2335, IEEE Computer Society. 124, 126

Saputra, M.R.U., Markham, A. & Trigoni, N. (2018). Visual slam and structure from motion in dynamic environments: A survey. *ACM Comput. Surv.*, **51**, 37:1–37:36. 3

Schultz, C. & Hermes, T. (2006). Digital keying methods. TZI-Bericht 40, Technologie-Zentrum Informatik, Bremen University. 96

Seitz, S.M., Matsushita, Y. & Kutulakos, K.N. (2005). A theory of inverse light transport. In *Proceedings of the International Conference on Computer Vision (ICCV)*. 72, 73

Shen, J., Yang, X., Jia, Y. & Li, X. (2011). Intrinsic images using optimization. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR)*, 3481–3487. 22, 25, 54, 59

Shen, J., Yan, X., Chen, L., Sun, H. & Li, X. (2014). Re-texturing by intrinsic video. *Information Sciences*, **281**, 726–735. 26

Shen, L. & Yeo, C. (2011). Intrinsic images decomposition using a local and global sparse representation of reflectance. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR)*, 697–704. 24, 27, 53, 54

Shen, L., Tan, P. & Lin, S. (2008). Intrinsic image decomposition with non-local texture cues. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR)*. 24

Shen, L., Yeo, C. & Hua, B.S. (2013). Intrinsic image decomposition using a sparse representation of reflectance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **35**, 2904–2915. 54

Shi, J., Dong, Y., Su, H. & Yu, S.X. (2017). Learning non-Lambertian object intrinsics across ShapeNet categories. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR)*, 5844–5853. 104

Shih, Y., Paris, S., Barnes, C., Freeman, W.T. & Durand, F. (2014). Style transfer for headshot portraits. *ACM Trans. Graph.*. 127

Shu, Z., Hadap, S., Shechtman, E., Sunkavalli, K., Paris, S. & Samaras, D. (2017). Portrait lighting transfer using a mass transport approach. *ACM Trans. Graph.*. 126, 136, 137, 143

Simonyan, K. & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *CoRR*, **abs/1409.1556**. 133

Sitzmann, V., Thies, J., Heide, F., Niessner, M., Wetzstein, G. & Zollhöfer, M. (2019a). Deepvoxels: Learning persistent 3d feature embeddings. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*. 147

Sitzmann, V., Zollhöfer, M. & Wetzstein, G. (2019b). Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *Advances in Neural Information Processing Systems*. 147

Steinbrcker, F., Sturm, J. & Cremers, D. (2014). Volumetric 3D mapping in real-time on a CPU. In *Proceedings of the International Conference on Robotics and Automation (ICRA)*. 55

Tan, J., Lien, J.M. & Gingold, Y. (2016). Decomposing images into layers via RGB-space geometry. *ACM Transactions on Graphics*, **36**, 7:1–14. 74

Tan, J., Echevarria, J. & Gingold, Y. (2018). Efficient palette-based decomposition and recoloring of images via RGBXY-space geometry. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)*, **37**, 262:1–10. 74

Tappen, M.F., Freeman, W.T. & Adelson, E.H. (2005). Recovering intrinsic images from a single image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **27**, 1459–1472. 22, 24, 25, 53

Theobalt, C., Ahmed, N., Lensch, H.P.A., Magnor, M.A. & Seidel, H. (2007). Seeing people in different light-joint shape, motion, and reflectance capture. *IEEE TVCG*, **13**, 663–674. 125

Thies, J., Zollhoefer, M., Stamminger, M., Theobalt, C. & Niessner, M. (2016). Face2face: Real-time face capture and reenactment of rgb videos. In *Proc. CVPR*. 126

Valentin, J., Vineet, V., Cheng, M.M., Kim, D., Shotton, J., Kohli, P., Niessner, M., Criminisi, A., Izadi, S. & Torr, P. (2015). SemanticPaint: Interactive 3D labeling and learning at your fingertips. *ACM Transactions on Graphics*, **34**, 154:1–17. 55, 58, 70

Vezhnevets, V., Sazonov, V. & Andreeva, A. (2003). A survey on pixel-based skin color detection techniques. In *Graphicon*. 58

Wang, T.Y., Ritschel, T. & Mitra, N.J. (2018). Joint material and illumination estimation from photo sets in the wild. *Computer Graphics Forum*, to appear. 106

Weber, D., Bender, J., Schnoes, M., Stork, A. & Fellner, D. (2013). Efficient GPU data structures and methods to solve sparse linear systems in dynamics applications. *Computer Graphics Forum*, **32**, 16–26. 36

Weber, H., Prévost, D. & Lalonde, J. (2018). Learning to estimate indoor lighting from 3d objects. *CoRR*, **abs/1806.03994**. 147

Weiss, Y. (2001). Deriving intrinsic images from image sequences. In *Proceedings of the International Conference on Computer Vision (ICCV)*, vol. 2, 68. 25, 54

Wen, Z., Liu, Z. & Huang, T.S. (2003). Face relighting with radiance environment maps. In *CVPR*. 126

Wenger, A., Gardner, A., Tchou, C., Unger, J., Hawkins, T. & Debevec, P. (2005). Performance relighting and reflectance transformation with time-multiplexed illumination. In *ACM SIGGRAPH 2005 Papers*, SIGGRAPH '05. 126, 127

Winnemller, H., Olsen, S.C. & Gooch, B. (2006). Real-time video abstraction. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, **25**, 1221–1226. 48

Wu, C., Zollhöfer, M., Niessner, M., Stamminger, M., Izadi, S. & Theobalt, C. (2014). Real-time shading-based refinement for consumer depth cameras. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)*, **33**, 200:1–10. 26, 34, 36, 37, 85

Wu, H. & Zhou, K. (2015). AppFusion: Interactive appearance acquisition using a Kinect sensor. *Computer Graphics Forum*, **34**, 289–298. 106

Wu, H., Wang, Z. & Zhou, K. (2016). Simultaneous localization and appearance estimation with a consumer RGB-D camera. *IEEE Transactions on Visualization and Computer Graphics*, **22**, 2012–2023. 11, 104, 106

Xia, R., Dong, Y., Peers, P. & Tong, X. (2016). Recovering shape and spatially-varying surface reflectance under unknown illumination. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)*, **35**, 187:1–12. 105

Xu, Z., Sunkavalli, K., Hadap, S. & Ramamoorthi, R. (2018). Deep image-based relighting from optimal sparse samples. *ACM Trans. on Graphics*. 124, 127, 147

Yamaguchi, S., Saito, S., Nagano, K., Zhao, Y., Chen, W., Olszewski, K., Morishima, S. & Li, H. (2018). High-fidelity facial reflectance and geometry inference from an unconstrained image. *ACM Trans. Graph.*, **37**. 124, 126, 143

Ye, G., Garces, E., Liu, Y., Dai, Q. & Gutierrez, D. (2014). Intrinsic video and applications. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, **33**, 80:1–11. 22, 25, 26, 39, 40, 46, 52, 54, 59, 65, 66, 67

Yu, Y. & Smith, W.A.P. (2018). Inverserendernet: Learning single image inverse rendering. *CoRR*, **abs/1811.12328**. 147

Yu, Y., Debevec, P., Malik, J. & Hawkins, T. (1999). Inverse global illumination: recovering reflectance models of real scenes from photographs. In *Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, 215–224. 11, 74, 105

Zeng, M., Zhao, F., Zheng, J. & Liu, X. (2013). Octree-based fusion for realtime 3D reconstruction. *Graphical Models*, **75**, 126–136. 55

Zhang, M., Chai, M., Wu, H., Yang, H. & Zhou, K. (2017). A data-driven approach to four-view image-based hair modeling. *ACM ToG*, **36**. 126

Zhang, R., Isola, P., Efros, A.A., Shechtman, E. & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. *CVPR*. 133

Zhao, H.K. (1996). *Generalized Schwarz Alternating Procedure for Domain Decomposition*. University of California, Los Angeles. 37

Zhao, Q., Tan, P., Dai, Q., Shen, L., Wu, E. & Lin, S. (2012). A closed-form solution to Retinex with nonlocal texture constraints. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **34**, 1437–1444. 22, 24, 54

Zheng, N., Jiang, Y. & Huang, D. (2019). Strokenet: A neural painting environment. In *International Conference on Learning Representations*. 148

Zhou, Q.Y. & Koltun, V. (2013). Dense scene reconstruction with points of interest. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, **32**, 112:1–8. 55

Zhou, T., Krhenbhl, P. & Efros, A. (2015). Learning data-driven reflectance priors for intrinsic image decomposition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 3469–3477. 22, 25, 54

Zhu, J.Y., Park, T., Isola, P. & Efros, A.A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*. 131

Zollhöfer, M., Stotko, P., Görlitz, A., Theobalt, C., Niessner, M., Klein, R. & Kolb, A. (2018). State of the art on 3d reconstruction with rgb-d cameras. *Computer Graphics Forum (Eurographics State of the Art Reports)*, **37**. 3, 11

Zollhfer, M., Niener, M., Izadi, S., Rhemann, C., Zach, C., Fisher, M., Wu, C., Fitzgibbon, A., Loop, C., Theobalt, C. & Stamminger, M. (2014). Real-time non-rigid reconstruction using an RGB-D camera. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, **33**, 156:1–12. 26, 34, 36, 85

Zollhfer, M., Dai, A., Innmann, M., Wu, C., Stamminger, M., Theobalt, C. & Niener, M. (2015). Shading-based refinement on volumetric signed distance functions. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, **34**, 96:1–14. 26, 34, 36, 37, 63, 105

Zoran, D., Isola, P., Krishnan, D. & Freeman, W.T. (2015). Learning ordinal relationships for mid-level vision. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 388–396. 25, 54