DISSERTATION

submitted to the

Combined Faculty of Natural Sciences and Mathematics

of the Ruprecht Karl University of Heidelberg, Germany

for the degree of

Doctor of Natural Sciences

Put forward by

M.Sc. Haebom Lee

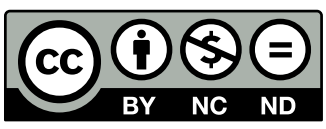
Born in: Busan, Republic of Korea

Oral examination:

Lighting Estimation in Outdoor Scenes

Advisor: Prof. Dr. Carsten Rother

Prof. Dr. Jean-François Lalonde



Abstract

This dissertation, titled “Lighting Estimation in Outdoor Scenes”, explores the vital aspect of light in computer vision, with a focus on the dynamic and complex nature of outdoor lighting. The work is motivated by the challenges in accurately capturing and interpreting outdoor lighting conditions, which are critical for applications in augmented reality (AR), 3D reconstruction, and autonomous driving technologies. Traditional single-image lighting estimation approaches often fall short due to issues like noise, inconsistency, and the complex interplay of natural elements. This thesis proposes new methodologies that extend beyond these limitations by incorporating both spatial and temporal analyses of lighting. This holistic approach allows for a more accurate and realistic interpretation of outdoor scenes, aiming to improve the realism of virtual objects in AR and the accuracy of various computer vision tasks.

The dissertation makes two major contributions: First, it explores the combination of intrinsic image decomposition and lighting estimation through a U-Net architecture, aiming to dissect images into albedo and shading components. While this exploration did not yield publication-worthy results, it provided valuable insights for future research. Second, it introduces advanced spatio-temporal outdoor lighting estimation methodologies, including a four-stage method and an end-to-end model utilizing a Transformer architecture for robust global sun direction estimation. These contributions signify an advancement in lighting estimation, with implications for various real-world applications.

Zusammenfassung

Diese Dissertation mit dem Titel “Lighting Estimation in Outdoor Scenes” untersucht den wichtigen Aspekt des Lichts in der Computer Vision, mit einem Fokus auf die dynamische und komplexe Natur in natürlichen Szenen im Frien. Die Arbeit ist motiviert durch die Herausforderungen bei der genauen Erfassung und Interpretation von Lichtverhältnissen im Außenbereich, die für Anwendungen in den Bereichen Augmented Reality (AR), 3D-Rekonstruktion und autonomes Fahren entscheidend sind. Herkömmliche Ansätze zur Schätzung der Beleuchtungsverhältnisse in Einzelbildern sind aufgrund von Problemen wie Rauschen, Inkonsistenz und dem komplexen Zusammenspiel natürlicher Elemente oft unzureichend. In dieser Arbeit werden neue Methoden vorgeschlagen, die über diese Einschränkungen hinausgehen, indem sie sowohl räumliche als auch zeitliche Analysen der Beleuchtung einbeziehen. Dieser ganzheitliche Ansatz ermöglicht eine genauere und realistischere Interpretation von Außenszenen, um den Realismus virtueller Objekte in AR und die Genauigkeit verschiedener Computer-Vision-Aufgaben zu verbessern.

Die Dissertation liefert zwei wichtige Beiträge: Erstens erforscht sie die Kombination von intrinsischer Bildzerlegung und Beleuchtungsschätzung durch eine U-Netz-Architektur, die darauf abzielt, Bilder in Albedo- und Schattierungskomponenten zu zerlegen. Obwohl diese Untersuchung keine veröffentlichungswürdigen Ergebnisse lieferte, lieferte sie wertvolle Erkenntnisse für die zukünftige Forschung. Zweitens werden fortschrittliche Methoden zur räumlich-zeitlichen Schätzung der Außenbeleuchtung vorgestellt, darunter eine vierstufige Methode und ein End-to-End-Modell, das eine Transformer-Architektur für eine robuste globale Schätzung der Sonnenrichtung nutzt. Diese Beiträge stellen einen Fortschritt in der Beleuchtungsabschätzung dar und haben Auswirkungen auf verschiedene reale Anwendungen.

Acknowledgements

Undertaking a PhD has been the most significant challenge of my life. I extend my deepest thanks to Prof. Dr. Carsten Rother for being an insightful advisor throughout this journey. He not only taught me the art of conducting research wisely but also continually inspired me with his relentless spirit. My sincere gratitude also goes to Prof. Dr. Jean-François Lalonde for graciously agreeing to be my second advisor. His work laid the groundwork for this dissertation.

This dissertation was developed during my time at Bosch Hildesheim, where I had the privilege of working alongside exceptional colleagues. I am particularly thankful for the consistent support from Prof. Dr. Jan Rexilius and Dr. Robert Herzog, whom I affectionately consider my Doktoronkel. I also wish to express my appreciation to my Doktorfreund, Christian Homeyer. Starting our PhDs together at the same university, we have shared numerous enriching discussions.

I cannot forget the people I met in Saarbrücken. Prof. Dr. Piotr Didyk, my first supervisor upon arriving in Germany, played a crucial role in helping me adjust to a new culture. Equally memorable is my (Junger) Doktoropa, Patric Gehl, the best landlord one could ask for, who also has a background in computer science dating back to the era of punch cards.

Turning to my roots in the Republic of Korea, I would like to first thank Dr. Jaehwan Kim for being considerate of my situation and providing great opportunities. I owe much of my foundational knowledge in Computer Graphics to Prof. Dr. Min H. Kim, who guided me during my Master's degree. I am also grateful to Kyoungjin Chi for opening the door to the field of computer graphics for me.

My family deserves my utmost appreciation. To my mother and sister, thank you for your solid belief in me and for encouraging me. I am also grateful to my parents-in-law and brother-in-law for their support. Without the combined support and encouragement from all of you, this journey would not have been possible. The path was filled with obstacles, but with the unwavering backing from my wife and daughter, Yoonah and Yeoleum, happiness was always within reach. I love you both so much!

Lastly, I dedicate this dissertation to my father, hoping he is reading this with a smile from heaven.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

The quest to replicate human vision in computer systems has been a cornerstone of technological advancement, leading to the burgeoning field of computer vision. At the heart of this discipline lies an elemental factor: light. Light not only illuminates our world but also fundamentally shapes our perception of it. In the realm of computer vision, understanding and harnessing the role of light is crucial for interpreting and reconstructing the visual environment as perceived by human eyes.

This dissertation, titled "Lighting Estimation in Outdoor Scenes", ventures into this intricate domain. The motivation for this exploration is driven by the challenges that arise in accurately capturing and interpreting the dynamic and complex nature of outdoor lighting. Traditional approaches to lighting estimation, particularly those limited to single-image analysis, often struggle with issues such as noise, inconsistency, and lack of temporal coherence. These challenges become particularly pronounced in outdoor scenes, where the interplay of natural light, shadows, and varying weather conditions adds layers of complexity to the task of lighting estimation.

The accurate estimation of lighting conditions in outdoor environments is more than an academic exercise; it holds substantial practical significance. Applications such as augmented reality (AR) heavily rely on precise lighting information to seamlessly blend virtual objects with real-world environments. In AR, the realism of virtual objects is directly linked to the accuracy of the lighting conditions under which they are rendered. Furthermore, a comprehensive understanding of outdoor lighting is essential for tasks such as 3D reconstruction, intrinsic image decomposition, material estimation, and shadow detection. These applications are central to various domains, including cinematography, architectural visualization, and autonomous driving technologies.

Confronting these challenges, this dissertation proposes innovative methodologies that

extend beyond the confines of traditional single-image analysis. It underscores the importance of considering both spatial and temporal aspects of lighting, recognizing that outdoor lighting, within a short time frame such as a couple of minutes, is essentially static yet observed from various locations and angles. This perspective allows for a more comprehensive analysis of lighting conditions across different spatial viewpoints and temporal moments, under the assumption of consistent lighting. By adopting this holistic approach, this dissertation aims to bridge the gap between the current state of lighting estimation and the demands of real-world applications, paving the way for more accurate and realistic interpretations of outdoor scenes.

1.2 Contributions

This dissertation presents two major contributions to the field:

Evaluating the Combination of Intrinsic Image Decomposition and Lighting Estimation: The research was to attempt to explore the decomposition of images into their intrinsic components, with a focus on simultaneous estimation of lighting conditions. The study primarily utilized a U-Net architecture to dissect images into albedo and shading, aiming to reveal insights into the intricate interplay of light and materials within a scene. Although the results were not compelling enough for publication, this exploratory endeavor provided valuable lessons and directions for future research in the field.

Spatio-Temporal Outdoor Lighting Estimation: A significant leap forward in lighting estimation, this work introduces two methodologies that overcome the limitations of prior single-image approaches. The first is a four-stage method that robustly estimates global sun direction by sampling across both spatial and temporal domains, significantly reducing noise and detecting outliers. The second is an end-to-end model employing a Transformer architecture, streamlining the estimation process and enhancing the realism of lighting models. These methods not only mark an advancement in lighting estimation but also have broad implications for applications in augmented reality and scene understanding.

1.3 List of Articles on which the Thesis Builds Upon

The remaining chapters of this thesis build upon the following two publications.

1. **Spatiotemporal Outdoor Lighting Aggregation on Image Sequences**
   Haebom Lee, Robert Herzog, Jan Rexilius, Carsten Rother
   DAGM German Conference on Pattern Recognition (GCPR) 2021

2. **Spatio-Temporal Outdoor Lighting Aggregation on Image Sequences Using Transformer Networks**
   Haebom Lee, Christian Homeyer, Robert Herzog, Jan Rexilius, Carsten Rother
   International Journal of Computer Vision (IJCV) 2023

1.4 Outline

This dissertation is structured as follows:

- **Chapter 1: Introduction**
  This chapter presents an overview of the dissertation, highlighting the motivation, challenges, and contributions of the research. It sets the stage for the subsequent chapters by outlining the key themes and objectives.

- **Chapter 2: Background**
  This chapter provides foundational knowledge in computer graphics, image processing, and deep learning. It contextualizes the research within the broader field, offering a perspective on the evolution and current state of these disciplines.

- **Chapter 3: Intrinsic Image Decomposition with Lighting Estimation**
  Focusing on intrinsic image decomposition, this chapter delves into the initial phase of the research. It explores the use of deep learning methods, particularly U-Net architecture and Siamese networks, to separate images into albedo and shading components and estimate lighting conditions at the same time.

- **Chapter 4: Spatio-temporal Outdoor Lighting Estimation**
  This chapter introduces advanced methodologies for outdoor lighting estimation. It first discusses a four-stage based approach that combines a single image-based method with statistical post-processing for spatio-temporal lighting estimation. The disadvantages of the first approach are resolved in the end-to-end model using a Transformer architecture, highlighting improvements in efficiency and realism.

- **Chapter 5: Conclusion**
  The final chapter synthesizes the findings of the dissertation, discussing the implications and potential applications of the research. It reflects on the journey of the study, acknowledging its limitations and proposing directions for future research.

Chapter 2

Background

In this chapter, we lay the foundational knowledge crucial for navigating the complex interplay between computer graphics, computer vision, and deep learning, as presented in this thesis. Our journey begins with an exploration of computer graphics, focusing on the principles of photorealistic rendering, the significance of accurate geometric modeling, and the intricacies of light models. These elements are pivotal for creating visually compelling digital imagery by simulating realistic interactions of light with various materials and surfaces. We then transition to a key concept in computer vision, Structure from Motion (SfM), which elucidates the process of reconstructing three-dimensional structures from sequences of two-dimensional images, a technique fundamental for understanding scene geometry and dynamics. Then we move on to the domain of deep learning, highlighting its transformative role in both computer vision and graphics. This section introduces neural network architectures, including Convolutional Neural Networks (CNNs) and U-Nets, underscoring their applications in tasks such as image classification, segmentation, and decomposition. By synthesizing these discussions, we aim to provide a comprehensive backdrop for the research presented in subsequent chapters, setting the stage for a deeper understanding of how these diverse yet interconnected fields contribute to advancements in digital image analysis and synthesis.

2.1 Computer Graphics and Vision

2.1.1 Rendering Equation

The Rendering Equation, introduced by James T. Kajiya in 1986, is a fundamental concept in computer graphics, offering a mathematical framework for simulating the way light interacts within a scene to achieve photorealistic rendering [8]. This equation models the distribution of light, capturing the essence of how light is emitted, scattered, absorbed, and reflected by

surfaces.

Expressed formally, the equation is:

$$L_o(p, \omega_o) = L_e(p, \omega_o) + \int_\Omega L_i(p, \omega_i) f_r(p, \omega_i, \omega_o)(\omega_i \cdot n) d\omega_i, \tag{2.1}$$

where $L_o(p, \omega_o)$ represents the outgoing radiance at point p in direction ω_o, and $L_e(p, \omega_o)$ is the radiance emitted by the surface at point p. The integral over the hemisphere Ω encompasses the incoming light $L_i(p, \omega_i)$ from all directions, with $f_r(p, \omega_i, \omega_o)$ as the bidirectional reflectance distribution function (BRDF), indicating how light is reflected at the surface. The term $(\omega_i \cdot n)$ quantifies the influence of the angle between the incident light direction ω_i and the normal to the surface n [9].

This equation is crucial for simulating direct and indirect lighting effects, underpinning the development of rendering techniques that aim for photorealism. Direct lighting considers light from a source reaching a surface directly, while indirect lighting accounts for light reflecting off surfaces before reaching the observer, contributing to the visual phenomena like shadows and reflections.

Key computational approaches to approximating solutions to the Rendering Equation include ray tracing and radiosity. Ray tracing is renowned for its ability to simulate complex optical effects, such as reflections and refractions, by tracing the paths of light rays through a scene [10]. Radiosity, alternatively, excels in modeling diffuse inter-reflections, capturing the soft illumination that arises when light bounces off surfaces, enhancing the overall realism of the scene [11].

While global illumination is a broader topic that will be explored in detail later, it is important to note here that the Rendering Equation lays the foundational principles for understanding these comprehensive lighting models. These principles guide the simulation of both direct and indirect lighting interactions within a scene, setting the stage for more advanced discussions on global illumination techniques.

The Rendering Equation also impacts the field of computer vision, where understanding light-surface interactions is essential for interpreting visual information. This understanding aids in reconstructing scenes from images, estimating material properties, and determining lighting conditions, demonstrating the equation's cross-disciplinary relevance [12, 13].

2.1.2 Camera Models

Camera models are essential in computer graphics and computer vision for simulating the process by which cameras capture light and form images. These models vary from simple abstractions that capture the essence of image formation to complex systems that accurately mimic real-world camera behaviors.

The simplest and most foundational camera model is the pinhole camera model, which serves as an idealized representation of how light travels from the scene to the image plane. The pinhole camera model is described by:

$$x = f\frac{X}{Z} \quad \text{and} \quad y = f\frac{Y}{Z}, \tag{2.2}$$

where x and y are the coordinates on the image plane, X, Y, and Z are the coordinates of a point in the scene, and f is the focal length of the pinhole camera. This model, despite its simplicity, forms the basis for understanding more complex camera behaviors and image formation processes [14].

Beyond the pinhole model, lens-based camera models introduce the effects of lenses, such as focus and depth of field, to simulate more accurately how cameras capture images. These models account for optical phenomena like lens distortion, which can cause straight lines in the scene to appear curved in the image. Lens distortion is typically modeled with radial and tangential components and corrected in computer vision applications using calibration techniques [15].

Another important aspect of camera models is the perspective projection, which captures how objects appear smaller as they are farther from the camera. Perspective projection is crucial for creating realistic three-dimensional effects in images and is represented by the equation:

$$\begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \frac{1}{Z} \begin{pmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}, \tag{2.3}$$

where the matrix represents the camera's intrinsic parameters that define its internal characteristics, including the focal length f and the principal point [16].

For applications that require modeling the camera's motion and orientation, the extrinsic parameters of the camera are also considered. These parameters define the position and orientation of the camera in the world and are essential for tasks such as 3D reconstruction and motion tracking. The relationship between the world coordinates and the camera coordinates is given by a rotation matrix R and a translation vector t, encapsulating the camera's viewpoint [16].

More advanced camera models also consider the effects of varying lighting conditions, sensor noise, and dynamic range, which influence how images are captured and processed. These factors are particularly relevant in the context of high dynamic range (HDR) imaging and low-light photography, where the limitations of camera sensors and the need for post-processing techniques become apparent [17].

Camera models play a pivotal role in both the generation of computer graphics and the analysis of images in computer vision. By understanding the principles of these models,

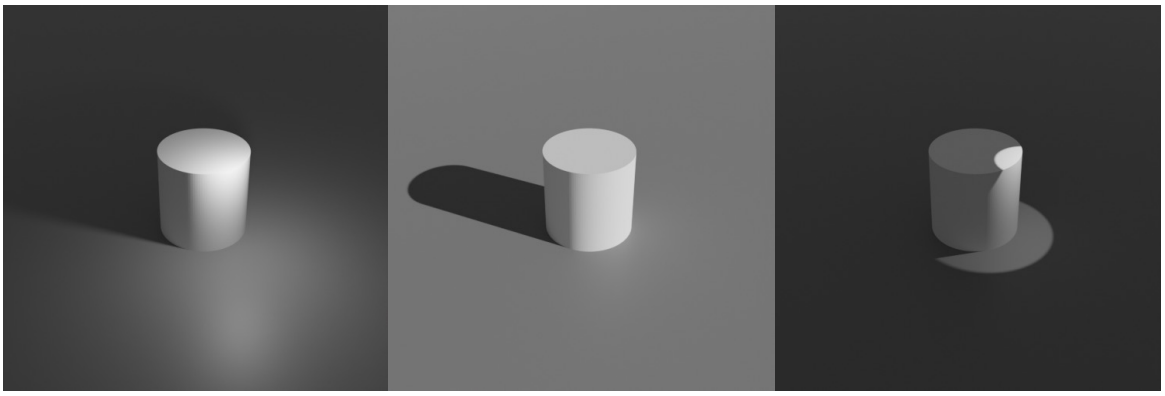


Figure 2.1. Illustration of various light models.

researchers and practitioners can simulate realistic camera effects, correct for optical distortions, and reconstruct three-dimensional scenes from two-dimensional images.

2.1.3 Light Models

Models for lights are one component for rendering images, and one aspect of the rendering equation. These models encapsulate the behavior of light, including its distribution, color and intensity.

Popular models for light sources in computer graphics are point lights, directional lights, ambient light, area light and spotlights. Point lights emit light uniformly in all directions from a single location, resembling bulbs or small light sources. Directional lights simulate sunlight or other distant light sources, casting parallel rays across the entire scene. Spotlights produce a conical beam of light, similar to a flashlight or stage lighting, allowing for focused illumination with controlled falloff [18]. Figure 2.1 shows examples of these light models.

This thesis addresses the task of light estimation in outdoor settings. In two works (chapter 3 and 4.2.1) we use a very simple outdoor model, where only the sun direction is modeled and estimated. This assumes that we only deal with images where the sun is actually shining and not covered by clouds. In order to deal with all weather conditions, we have also developed an approach which works for all outdoor lighting conditions. For this we use the so-called Lalonde-Matthews outdoor illumination model [19], which is defined in detail in section 4.2.2. Briefly, the model has two components. One models the shape, color, and position of the sun. The other deals with the remining aspects of the sky and models the color and turbidity of the sky.

Figure 2.2. Illustration of various materials.

2.1.4 Material

Materials play a crucial role in computer graphics by defining the appearance of objects under different lighting conditions. The interaction of light with materials is determined by properties such as color, texture, reflectivity, transparency, and surface roughness. These properties influence how materials absorb, reflect, and transmit light, contributing to the overall realism of a scene. Figure 2.2 shows several spheres made of different materials.

The Bidirectional Reflectance Distribution Function (BRDF) is a fundamental tool for modeling the reflective properties of materials. It describes the relationship between incoming light and the light reflected by a surface, depending on the viewing and illumination directions. The BRDF is crucial for simulating how different materials respond to light, allowing for the creation of a wide range of appearances, from matte surfaces with diffuse reflection to shiny surfaces with specular highlights [20].

In addition to reflectance, materials may also exhibit translucency or transparency, characterized by their ability to transmit light. The Bidirectional Transmission Distribution Function (BTDF) models this behavior, accounting for light that passes through a material and emerges in different directions. Together, the BRDF and BTDF enable the simulation of complex materials like glass, water, and thin fabrics [21].

Textures add another layer of realism to material appearance, representing surface details such as patterns, bumps, and scratches. Texture mapping applies image textures to 3D models, while bump mapping and normal mapping simulate surface irregularities by

altering the surface normals used in lighting calculations. These techniques enhance the perceived depth and detail of materials without significantly increasing the complexity of the 3D models [22].

Materials also have intrinsic properties such as subsurface scattering, which occurs when light penetrates the surface of a translucent material, scatters internally, and exits at different points. This effect is essential for rendering materials like skin, marble, and wax, where light diffusion contributes to the material's soft, glowy appearance [23].

The accurate simulation of material properties is essential for creating visually compelling images in computer graphics. By leveraging models like the BRDF, BTDF, and physically based rendering (detailed in 2.1.6) principles, along with techniques for texture and detail simulation, graphics artists and researchers can replicate the diverse and nuanced ways materials interact with light in the real world.

In most parts of the thesis, in particular chapter 4, we do not consider the task of estimating material properties, such as parts of a BRDF. In chapter 3 we utilize a very simple model for material. This means that we aim at estimating the true surface colors in RGB format. These colors are independent of any lighting effects, such as shading and shadows. To decompose an image into two images, one modeling the surface colors the other one the lighting effects, is called intrinsic image decomposition in computer vision, and explained in detail in chapter 3.

2.1.5 Geometry

Geometry is the third important concept in computer graphics and computer vision, providing the framework within which light interactions are simulated to create realistic images. The geometric shape and structure of objects determine how they interact with light, influencing the appearance of shadows, reflections, and refractions.

In computer graphics, geometric models are used to represent the shape of objects in a scene. The most basic models are primitives, such as spheres, cubes, and cylinders, which can be combined or modified to create more complex forms. For greater complexity and detail, polygonal meshes, composed of vertices, edges, and faces, are widely used. These meshes can accurately represent intricate shapes by adjusting the density and arrangement of polygons, though this increases computational complexity [18].

Surface representation is another critical aspect of geometry, involving the detailed modeling of object surfaces to capture textures, bumps, and other features that affect light interaction. Techniques such as bump mapping and displacement mapping are employed to simulate surface irregularities without the need for high-polygon models, enhancing the realism of rendered images with minimal impact on computational resources [24].

Parametric and implicit surfaces offer alternative methods for defining complex geometries. Parametric surfaces, such as Bézier surfaces and NURBS (Non-Uniform Rational B-Splines), are defined by mathematical functions that provide precise control over shape and smoothness. Implicit surfaces are defined by a scalar field without explicit edges or vertices, enabling the modeling of soft transitions and organic forms [25].

In the realm of computer vision, geometry plays a crucial role in interpreting the three-dimensional structure of scenes from two-dimensional images. Techniques such as structure from motion (detailed in 2.1.7) and stereo vision rely on geometric principles to estimate the spatial arrangement of objects, their shapes, and their positions relative to the camera. These methods are fundamental for tasks such as 3D reconstruction, object recognition, and navigation in robotics [14].

The interaction between geometry and light is a key factor in achieving photorealism in computer-generated imagery. Shadow casting, an essential element of realistic scenes, is directly influenced by the geometric form of objects. Advanced rendering techniques, including ray tracing and global illumination models, simulate light behavior as it intersects with geometric forms, calculating shadows, reflections, and refractions based on the shape and orientation of objects [9].

Geometry not only defines the visible structure of a scene but also influences the distribution and appearance of light and shadow, contributing to the perception of depth, scale, and material properties. By leveraging sophisticated geometric models and computational techniques, computer graphics and computer vision can create highly realistic simulations of physical environments, bridging the gap between digital imagery and the complexities of the real world.

2.1.6 Illumination Models

Two primary lighting models are defined in the study of illumination: local illumination, also termed object-oriented lighting, and global illumination. The distinction lies in the fact that with local illumination each object is treated separately. In contrast, global illumination encompasses the comprehensive dynamics of light as it scatters and reflects across multiple objects within a scene, thus providing a more accurate simulation of light's interaction with its environment.

One of the primary models used to represent light in computer graphics is the Phong lighting model, which includes three components: ambient, diffuse, and specular lighting (see Figure 2.3). The ambient light represents a constant light present in the scene to simulate the effect of indirect light bouncing off surfaces. Diffuse lighting models the way light scatters in many directions when it hits a rough surface, making it appear uniformly

Figure 2.3. Illustration of the Phong lighting model, showcasing ambient, diffuse, and specular components.

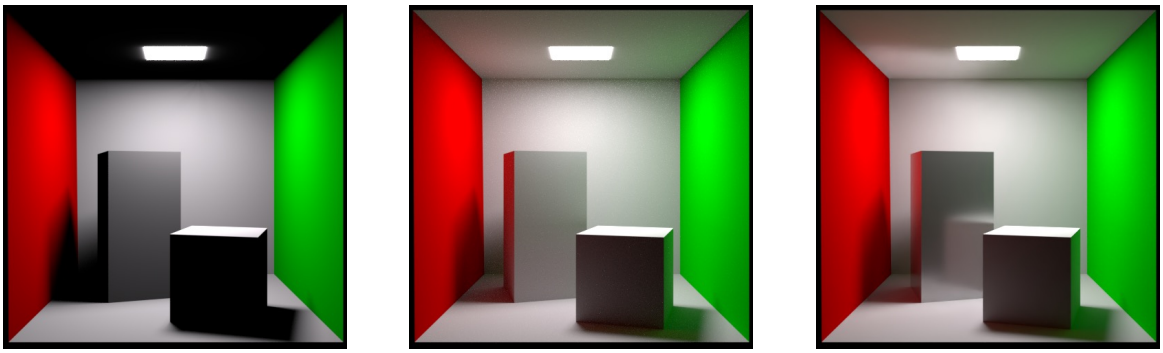


Figure 2.4. Illustration of different lighting models and techniques: direct lighting as immediate light source interaction, path tracing for simulating light paths and indirect illumination, and physically based rendering for realistic material and light interactions.

illuminated from all angles. Specular lighting captures the bright spots that appear on shiny surfaces when viewed from specific angles, contributing to the perception of glossiness [26].

Global Illumination (GI) encompasses a set of techniques in computer graphics aimed at simulating the complex interactions of light in a scene [27]. Unlike local illumination models, which only consider direct light from sources to surfaces, GI accounts for both the direct and indirect light contributions. This includes light bouncing off multiple surfaces, color bleeding, caustics, and the subtle diffusion of light through translucent materials (see Figure 2.4).

One of the foundational techniques for simulating global illumination is ray tracing. Enhanced ray tracing algorithms extend the basic concept by tracing multiple secondary rays at points of reflection, refraction, or transmission. This allows for the accurate simulation of effects such as soft shadows, depth of field, and indirect lighting. Despite its computational

intensity, ray tracing remains a gold standard for high-quality rendering due to its ability to produce highly realistic images [10].

Radiosity is another crucial method for calculating global illumination, focusing primarily on the diffuse inter-reflection of light between surfaces. Unlike ray tracing, radiosity solves the energy transfer equation for the entire scene as a global system, resulting in a solution that accurately reflects the color bleeding effect where the color of illuminated surfaces affects the light color on adjacent surfaces. Radiosity is particularly effective for scenes dominated by diffuse interactions and provides a solution that can be precomputed for static scenes [11].

Photon mapping, introduced by Henrik Wann Jensen, combines elements of both ray tracing and radiosity to efficiently simulate global illumination effects, including caustics. In this technique, photons are emitted from light sources, traced through the scene, and stored in a photon map. This map is then used to estimate the indirect illumination at various points in the scene, allowing for the efficient and scalable simulation of complex lighting effects [28].

More recent advancements in global illumination include techniques such as path tracing and light transport algorithms, which attempt to simulate the full path of light rays as they bounce through a scene. Path tracing, a Monte Carlo method, uniformly samples light paths connecting the camera and the light sources via scattering events in the scene. This technique, while computationally demanding, can produce highly realistic images with accurate global illumination effects over time [8].

Precomputed Radiance Transfer (PRT) offers a way to approximate global illumination in dynamic scenes under fixed lighting conditions. PRT precomputes how light interacts with surfaces and stores this information in textures or vertex attributes, allowing for real-time rendering of complex lighting effects, including soft shadows and inter-reflections, in interactive applications [29].

Recent advances in light modeling have focused on physically based rendering (PBR), which aims to simulate light behavior and material properties with high fidelity to physical laws. PBR frameworks utilize more sophisticated lighting models to achieve realistic shading and material appearance, considering factors like energy conservation and the Fresnel effect to enhance realism [30].

Global illumination significantly enhances the realism of computer-generated images by accurately simulating the nuanced and complex ways light interacts within a scene. As computational power increases and algorithms become more efficient, the incorporation of global illumination techniques in real-time rendering applications, such as video games and virtual reality, continues to grow, bridging the gap between real-time performance and photorealistic quality.

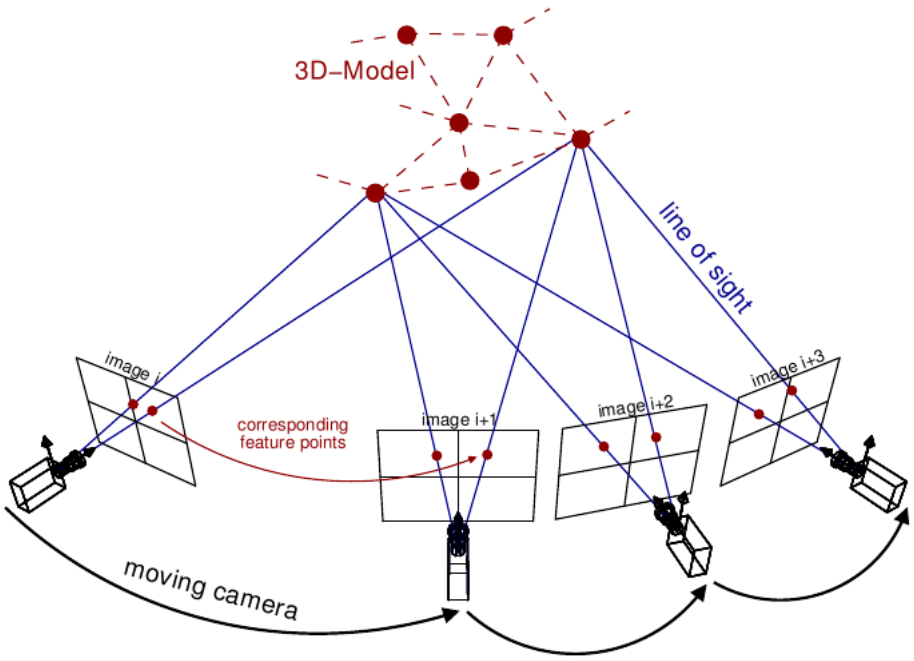


Figure 2.5. Structure from Motion (SfM) photogrammetric principle. This image was copied from the article [1]

Illumination models, whether local or global, do not influence our work in estimating lighting models. This is because our method does not incorporate an explicit illumination model (or rendering process), for example, within the loss function. Instead, our loss function directly compares the parameters of the estimated lighting model with those of the ground truth lighting model.

2.1.7 Structure from Motion (SfM)

Structure from Motion (SfM) is a process in computer vision that reconstructs the three-dimensional geometry of a scene from a series of two-dimensional images. By analyzing the apparent motion of objects across multiple images taken from different viewpoints, SfM algorithms can infer the spatial layout of the scene and the camera's path during image capture. This technique is foundational for applications in 3D reconstruction, aerial mapping, heritage preservation, and augmented reality.

The core principle behind SfM is the extraction of feature points across the set of images and the identification of correspondences between these features across views (see Fig. 2.5). Features are typically points of interest within the image that can be reliably detected and matched, such as corners, edges, or distinct texture patterns. Once correspondences are established, it is possible to estimate the relative camera poses (positions and orientations)

and the three-dimensional coordinates of the feature points in the scene [14].

SfM algorithms can be categorized into two main types: incremental (sequential) and global. Incremental SfM adds images one at a time to the reconstruction, iteratively updating the 3D model and camera poses with each new image. This method is intuitive and can handle large-scale problems but may suffer from error accumulation over the sequence. Global SfM, on the other hand, attempts to solve for all camera poses and 3D point positions simultaneously, often leading to more accurate and consistent reconstructions, especially for looped sequences where the camera returns to its starting point [31].

Camera calibration is a critical step in SfM, as it involves determining the intrinsic parameters of the camera (such as focal length and lens distortion) that affect image formation. Accurate calibration is essential for precise 3D reconstruction, and while some SfM approaches require pre-calibrated cameras, others can auto-calibrate based on the image data itself, estimating both the scene structure and the camera parameters simultaneously [32].

Bundle adjustment is the final optimization step in SfM, where the initial estimates of camera poses and 3D points are refined to minimize the reprojection error, which is the difference between the observed feature positions in the images and the projected positions from the 3D model. This non-linear least squares optimization is computationally intensive but crucial for achieving high-quality reconstructions [33].

SfM has enabled a wide range of applications, from creating 3D models of architectural sites to generating topographic maps from drone imagery. The ability to reconstruct accurate 3D structures from standard photographs offers a versatile tool for understanding and documenting the physical world in digital form. A notable example of advanced SfM software is COLMAP, which automates many aspects of the SfM and MVS processes, providing an end-to-end pipeline for 3D reconstruction from images. COLMAP's features include automatic image matching, robust reconstruction algorithms, and support for dense point cloud generation, making it a powerful solution for both academic research and practical applications in 3D modeling [34, 35]. In particular, SfM played a key role in chapter 4, where it was used to predict the camera yaw angle across image sequences.

2.2 Deep Learning

2.2.1 Introduction

Deep learning, a subset of machine learning, has emerged as a powerful tool in the analysis and interpretation of complex datasets, particularly in the field of computer vision. At its core, deep learning utilizes neural networks with multiple layers – hence the term "deep" – to model intricate patterns and relationships within data. This approach has revolutionized

the way computers interpret visual information, enabling significant advancements in tasks ranging from image recognition to semantic segmentation and beyond.

The relevance of deep learning in computer vision, especially concerning lighting analysis, cannot be overstated. In environments where lighting conditions vary extensively, traditional algorithms struggle to maintain consistency and accuracy. Deep learning models, however, excel in these scenarios by learning feature representations that are robust to changes in lighting, perspective, and background noise. This capability is particularly valuable in applications such as autonomous vehicles, where understanding the environment under different lighting conditions is crucial, and in augmented reality, where accurate lighting estimation enhances the realism of virtual objects overlaid on real-world scenes.

One of the foundational concepts in deep learning is the ability of models to learn hierarchical representations. In the context of computer vision, this means that lower layers of a neural network might learn to recognize edges and textures, while deeper layers can interpret more complex features such as shapes and objects. This hierarchical learning process is akin to the way human vision system operates, from basic perception to complex interpretation, making deep learning models particularly adept at understanding visual scenes.

Moreover, the advent of deep learning has facilitated the development of models that can learn directly from raw data, eliminating the need for manual feature extraction, which was a significant bottleneck in traditional machine learning approaches. This shift has not only streamlined the workflow for developing computer vision applications but has also opened up new possibilities for analyzing and understanding visual data in unprecedented detail.

As we delve deeper into the specifics of deep learning models such as Feedforward Neural Networks, Convolutional Neural Networks, Transformers, and U-Net architectures, we will explore their unique contributions to the field of computer vision. Each model offers distinct advantages for interpreting visual data, particularly in scenarios complicated by varying lighting conditions, showcasing the versatility and power of deep learning in pushing the boundaries of what is possible in image analysis and lighting estimation.

2.2.2 Feedforward Neural Networks

Feedforward Neural Networks (FNNs) are the most basic form of artificial neural network architecture, characterized by a unidirectional flow of information. This structure includes an input layer, several hidden layers, and an output layer, with connections between nodes that do not form cycles. The straightforward progression of data from input to output allows FNNs to model a wide array of functions, making them a versatile tool in machine learning and computer vision applications [36].

The strength of FNNs lies in their capacity to approximate any continuous function, a

property formally recognized as the universal approximation theorem. This theorem asserts that FNNs can capture a vast range of relationships within data, provided they have sufficient neurons in their hidden layers [37].

$$y = f(\mathbf{W} \cdot \mathbf{x} + \mathbf{b}) \tag{2.4}$$

This equation represents the basic operation within a neuron of a feedforward neural network. Here, y is the output, f is an activation function (such as sigmoid, ReLU, etc.), $\mathbf{W}$ is a weight matrix, $\mathbf{x}$ is the input vector, and $\mathbf{b}$ is a bias vector. This operation is fundamental to how FNNs model complex relationships and patterns in data.

FNNs find extensive use in tasks such as classification and regression, where the objective is to infer an output from a given set of inputs based on learned data representations. The model's input layer receives data (for instance, pixel values from images), which is then processed through one or more hidden layers that extract and learn patterns. The output layer generates the final prediction or classification result, utilizing the features identified by the network.

A significant challenge associated with FNNs is their handling of raw pixel data from large images, primarily due to the substantial number of input features and the absence of mechanisms to exploit spatial or temporal structures within the data. This issue is often mitigated by preprocessing the data into a more compact form or by integrating FNNs with architectures like convolutional neural networks (CNNs), which are inherently better suited for processing image data due to their convolutional layers that capture spatial hierarchies [38].

Despite such challenges, FNNs have remained foundational in the evolution of neural network architectures, providing a critical basis for the development of more complex models tailored to specific challenges in computer vision and beyond. Their simplicity, coupled with the robustness of their learning capabilities, continues to make FNNs an essential component of the machine learning ecosystem [39].

2.2.3 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) have revolutionized the field of computer vision by enabling the effective processing and interpretation of visual data. Distinct from Feedforward Neural Networks, CNNs are adept at automatically learning spatial hierarchies of features from image data, which positions them as a critical tool for a wide range of computer vision tasks [40].

The architecture of a CNN is distinguished by its composition of convolutional layers, pooling layers, and fully connected layers. Convolutional layers employ a set of learnable

filters that are applied to the input images, effectively capturing spatial features like edges, textures, and shapes. This mechanism maintains the spatial relationships between pixels, facilitating the network's ability to learn image features efficiently.

$$\mathbf{Z}^{[l]} = \mathbf{W}^{[l]} * \mathbf{A}^{[l-1]} + \mathbf{b}^{[l]} \tag{2.5}$$

This equation describes the computation within a convolutional layer, where $\mathbf{Z}^{[l]}$ represents the output feature map of the l^{th} layer, $\mathbf{W}^{[l]}$ denotes the weights (filters) of the l^{th} layer, $*$ signifies the convolution operation applied between the filters and the input volume $\mathbf{A}^{[l-1]}$ from the previous layer $(l-1)$, and $\mathbf{b}^{[l]}$ is the bias. The convolution operation here captures the spatial dependencies in the input through the application of filters, allowing the network to efficiently learn spatial hierarchies of features such as edges, textures, and shapes in the input images.

Pooling layers serve to reduce the dimensionality of the data, aggregating the outputs of neuron clusters at one layer into a single neuron in the subsequent layer. This reduction in parameters and computation helps in achieving feature detection that is invariant to scale and orientation, thus bolstering the network's generalization capabilities [41].

Fully connected layers, situated towards the end of the CNN architecture, consolidate the features learned by previous layers to produce the final output, such as a class label. The integration of these layers allows CNNs to comprehend the intricate relationships present within visual data, enabling applications ranging from image classification and object detection to comprehensive scene understanding.

CNNs have been pivotal in advancing the field of computer vision, enhancing the analysis and interpretation of visual content. Their development continues to explore new frontiers, contributing to innovative research and applications across a variety of domains. The adaptability and efficiency of CNNs in processing complex visual information have established them as a cornerstone of modern computer vision techniques.

2.2.4 Transformers

Transformers have significantly influenced the field of computer vision by introducing an innovative approach originally developed for natural language processing (NLP). Central to the Transformer architecture is the self-attention mechanism, which enables the model to assess and prioritize different segments of the input data, facilitating a comprehensive understanding of the context within images [42].

The self-attention mechanism is composed of queries (Q), keys (K), and values (V), derived from the input data to compute attention scores. These scores determine the emphasis placed on various parts of the input sequence:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.6}$$

In this equation, d_k denotes the dimensionality of the keys, facilitating the model's capability to dynamically allocate focus across the input data based on relevance to the given task.

The architecture encompasses an encoder and a decoder, each consisting of multiple layers. The encoder processes the input data, generating attention-based feature representations, which the decoder uses to produce the output sequence. This sequence could range from textual outputs in NLP applications to structured data forms pertinent to various computer vision tasks. The term "FFN" below stands for FeedForward Network, a component within both the encoder and decoder that consists of layers performing linear transformations followed by nonlinear activations, crucial for enhancing the model's ability to process complex patterns:

$$\text{Encoder: } \mathbf{Z} = \text{Attention}(Q, K, V) + \text{FFN}(\mathbf{Z}) \tag{2.7}$$

$$\text{Decoder: } \mathbf{Y} = \text{Attention}(\mathbf{Z}, \mathbf{Z}, \mathbf{Z}) + \text{FFN}(\mathbf{Y}) \tag{2.8}$$

Positional encoding is integral to the Transformer model, compensating for its intrinsic lack of sequential order understanding. Through the application of sine and cosine functions, positional encoding imparts the sequence order of inputs, crucial for processing image data effectively:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right) \tag{2.9}$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right) \tag{2.10}$$

The adaptation of the Transformer architecture for computer vision, as exemplified by the Vision Transformer (ViT) model, has demonstrated notable success. ViT processes images in patches, treating them as sequences to leverage the self-attention mechanism for capturing the scene's global context, proving effective in a variety of computer vision tasks [43].

The flexibility of Transformers, characterized by their dynamic focus and adaptability to various input data scales, positions them as a robust tool in the computer vision domain. Ongoing advancements suggest that integrating Transformers with conventional neural network architectures could further enhance their utility, opening new avenues for research and practical applications in the field.

2.2.5 U-Net

The U-Net architecture has become a cornerstone in the field of computer vision, acclaimed for its remarkable efficiency in image segmentation tasks. Originating from biomedical image segmentation, U-Net's design is optimized for precise localization, proving to be exceptionally effective across a wide range of imaging contexts where capturing detailed texture and contextual information is essential for accurate analysis [44].

Characterized by its distinctive U-shaped structure, U-Net comprises a contracting path that captures the context and a symmetric expanding path for precise localization. This dual-path architecture enables the model to deliver high performance even with limited training data, through the effective use of data augmentation techniques. The architecture is succinctly described as:

$$\text{U-Net} : \mathbf{I} \rightarrow \text{Encoder} \rightarrow \text{Decoder} \rightarrow \mathbf{O} \tag{2.11}$$

Here, **I** and **O** represent the input image and the output segmentation map, respectively. The encoder section adopts a convolutional network structure, utilizing convolutions and pooling operations to distill contextual information from the image. The decoder section, in contrast, employs transposed convolutions to expand the feature maps back to the original input size. Skip connections are implemented between mirrored layers in the encoder and decoder, enriching the decoder with high-resolution details and enabling more accurate image reconstruction.

$$\text{Skip Connection} : \mathbf{F}_{\text{encoder}} \oplus \mathbf{F}_{\text{decoder}} \tag{2.12}$$

In this equation, $\mathbf{F}_{\text{encoder}}$ and $\mathbf{F}_{\text{decoder}}$ denote feature maps from the encoder and decoder pathways, respectively, with $\oplus$ indicating a concatenation operation.

Beyond its origins in biomedical imaging, U-Net has demonstrated broad applicability across a variety of computer vision challenges, showcasing its adaptability and effectiveness in tasks beyond its initial scope. The architecture's ability to accurately segment images makes it a valuable tool for any application requiring detailed analysis of visual data.

Furthermore, the potential for U-Net to be integrated with other deep learning models, such as CNNs and Transformers, opens up new avenues for enhancing its performance. Such integrations can leverage U-Net's strength in precise localization alongside the broader contextual understanding offered by other architectures, setting the stage for innovative advancements in computer vision.

2.3 Summary

The exploration of "Computer Graphics and Vision" highlights the essential role of lighting in creating realistic digital images. Techniques like the rendering equation and global illumination models are key in simulating how light interacts with objects, influencing their appearance and how the scene is perceived. The rendering equation provides the basis for understanding light's behavior, leading to the development of advanced algorithms for photorealistic rendering. Meanwhile, global illumination techniques such as ray tracing and radiosity give detailed simulations of light's diffuse and specular interactions, improving the realism of computer-generated imagery. This background information is important for understanding the details in chapter 3, where we tackle intrinsic image decomposition along with lighting estimation.The section on structure from motion (SfM) shows the vital link between geometry and lighting, with the accurate reconstruction of three-dimensional scenes from two-dimensional images heavily dependent on recognizing light's impact on visual perception. Beyond its basic function in reconstructing 3D scenes, SfM proves invaluable in estimating camera trajectories, supporting the work in chapter 4, where determining the camera's yaw angle is necessary for conducting spatio-temporal lighting estimation.

In the "Deep Learning" section, the focus shifts to the transformative impact of neural network architectures on interpreting visual data, especially in the context of lighting. Feedforward Neural Networks (FNNs), while foundational, reveal limitations in handling spatial hierarchies, paving the way for the adoption of Convolutional Neural Networks (CNNs) in tasks requiring an understanding of spatial relationships and lighting conditions. CNNs, with their ability to learn hierarchical feature representations from image data, have significantly advanced the field's capability to process and analyze visual information under varied lighting. The introduction of Transformers and U-Net architectures further expands the toolbox available for computer vision tasks, offering novel approaches to model global context and achieve precise localization in image segmentation, respectively. These advancements underscore the synergy between deep learning models and traditional computer vision techniques, enhancing the ability to analyze and manipulate lighting in images for improved realism and accuracy. These neural networks are extensively utilized in chapters 3 and 4, with U-Net serving as the principal architecture in chapter 3, and Transformers contributing significantly to the end-to-end spatio-temporal lighting estimation network presented in chapter 4.

The integration of computer graphics and vision techniques with deep learning models represents a significant leap forward in creating and interpreting realistic digital imagery. By combining the principles of light modeling and geometry with the advanced pattern recognition capabilities of neural networks, researchers and practitioners can tackle com-

plex challenges in lighting analysis, from estimating lighting conditions to reconstructing photorealistic scenes. The ongoing evolution of these technologies promises to unlock new possibilities in digital imaging, augmented reality, and beyond, driving further innovation in how we capture, simulate, and understand the visual world.

Chapter 3

Lighting Estimation combined with Intrinsic Image Decomposition

The quest for understanding and replicating human vision in machines has led to significant advancements in computer vision, particularly in the ability to interpret and reconstruct the visual world as perceived by human eyes. This challenge encompasses the decomposition of images into their intrinsic components. This chapter introduces an exploratory phase of research focused on intrinsic image decomposition using deep learning methods, specifically targeting the challenge of lighting estimation within this context.

Intrinsic image decomposition, a fundamental problem in computer vision, aims to separate an image into its constituent elements, typically albedo and shading. Albedo represents the intrinsic color of objects, unaffected by lighting, while shading embodies the effects of illumination. This separation is crucial for numerous applications, including photorealistic rendering, image editing, and understanding scene dynamics.

The initial design incorporated a U-Net architecture for decomposing RGB images into albedo and shading components. The hypothesis was that the latent vector at the deepest level of the U-Net contained encoded information about the lighting conditions of the input image. An experimental attachment of a Multi-Layer Perceptron (MLP) network to the deepest layer of U-Net aimed to extract the direction of sunlight in the image.

The prototype of this approach was initially trained on synthetic images rendered using Blender [45], taking advantage of the access to ground truth intrinsic components available during the rendering process. However, the extension of our method to real-world images presented challenges due to the unavailability of ground truth intrinsic components for such scenes. To address this, we adapted our evaluation strategy for practical application to real images, employing alternative training methods that enable the network to decompose images into their intrinsic components without the need for direct ground truth comparison.

In adapting to real-world data limitations, we introduced a new approach utilizing triplet

loss within a Siamese network configuration. This configuration relies on the relationships between triplets of images: i) a reference image, ii) a second image with the same albedo but different shading, and iii) a third image with consistent shading but a different albedo. This approach enables an indirect assessment of model performance, overcoming the challenge of absent ground-truth intrinsic images for real-world applications.

Although the triplet loss approach eliminated the need for ground-truth intrinsic images, identifying such triplets in the real world posed a significant challenge. This resulted in a heavy reliance on synthetic scenes rendered in Blender. The network showed proficiency in decomposing simple synthetic images but struggled with more complex, realistic synthetic scenes.

This research phase, despite its challenges, offers valuable insights and learnings, demonstrating the iterative and often non-linear nature of scientific inquiry. This chapter documents this journey, providing a detailed account of the methods, experiments, and results, illustrating a narrative of learning and adaptation in scientific research.

The following sections discuss the related work that informed this research, the methods developed, the experimental setup and results, and the conclusions drawn from this exploratory phase.

3.1 Related Work

The exploration of intrinsic image decomposition in computer vision, particularly in the context of lighting estimation, has been significantly influenced by various foundational and innovative studies. This section contextualizes the research within the broader landscape, highlighting key contributions and methodologies that have shaped the field.

The concept of intrinsic images, introduced by Barrow et al. [46], established an important foundation for the field. Their work proposed the decomposition of images into two key components: illumination and reflectance. This approach has not only guided subsequent research in intrinsic image decomposition but also provided a conceptual framework for understanding how images can be analytically broken down into their fundamental elements. Their methodology has been the cornerstone of many later studies, influencing a wide range of applications in computer vision and image processing.

With the advent of deep learning, the field witnessed a transformative leap. In 2015, Narihira et al. [47] introduced a novel approach to intrinsic image decomposition using Convolutional Neural Networks (CNNs). This strategy, known as direct intrinsics, involved training a CNN to predict output albedo and shading channels directly from an input RGB image. This method represented a significant departure from traditional techniques, showcasing the robust potential of deep learning in handling complex tasks like intrinsic image

decomposition. It marked a shift towards more data-driven approaches, leveraging the capabilities of neural networks to interpret and process visual information in ways that were previously challenging or impossible with conventional methods.

The introduction of the U-Net architecture by Ronneberger et al. [44] further advanced the field. Originally developed for biomedical image segmentation, its unique architecture featured a contracting path to capture context and a symmetric expanding path for precise localization. This design proved to be highly effective in various image processing tasks, including intrinsic image decomposition. The U-Net's ability to handle fine details in images while maintaining contextual information made it an influential tool in the study of image decomposition, inspiring new ways to approach complex vision tasks.

In the realms of computer graphics and vision, lighting estimation is crucial for understanding visual perception and image formation. Haber et al. [48] made a notable contribution by developing a method to recover the reflectance of static scenes using images captured under various distant illuminations. This method, utilizing a wavelet-based relighting framework and accommodating illumination variations, significantly expanded the scope of lighting estimation applications. Building upon this, Lalonde et al. [49] and their later work [19] provided foundational techniques for estimating natural illumination from a single outdoor image, thus enhancing the realism in computer-generated imagery. These collective efforts have established a comprehensive framework for lighting estimation, essential for achieving realistic and context-aware processing in computer-generated imagery.

Finally, the introduction of the FaceNet system by Schroff et al. [50] highlighted the effectiveness of triplet loss in learning fine-grained image distinctions. Employing a deep convolutional network trained to optimize the embedding itself, their work demonstrated the potential of using triplet loss in deep learning models. This approach inspired the adaptation of triplet loss in intrinsic image decomposition, providing a new perspective on how to train models to recognize and differentiate between intricate image features.

The following sections will elaborate on the integration of these influences into the methodology, the experimental setup, and the insights derived from the results.

3.2 Method

At the core of the proposed network, the U-Net architecture, known for its efficacy in image-to-image translation tasks, was adapted to decompose RGB images into their albedo and shading components. Central to this approach was the hypothesis that the latent vector at the deepest level of U-Net contained encoded information about lighting conditions. To explore this hypothesis, an additional Multi-Layer Perceptron (MLP) network was integrated into the deepest layer of U-Net, aiming to extract sunlight direction information from the images.

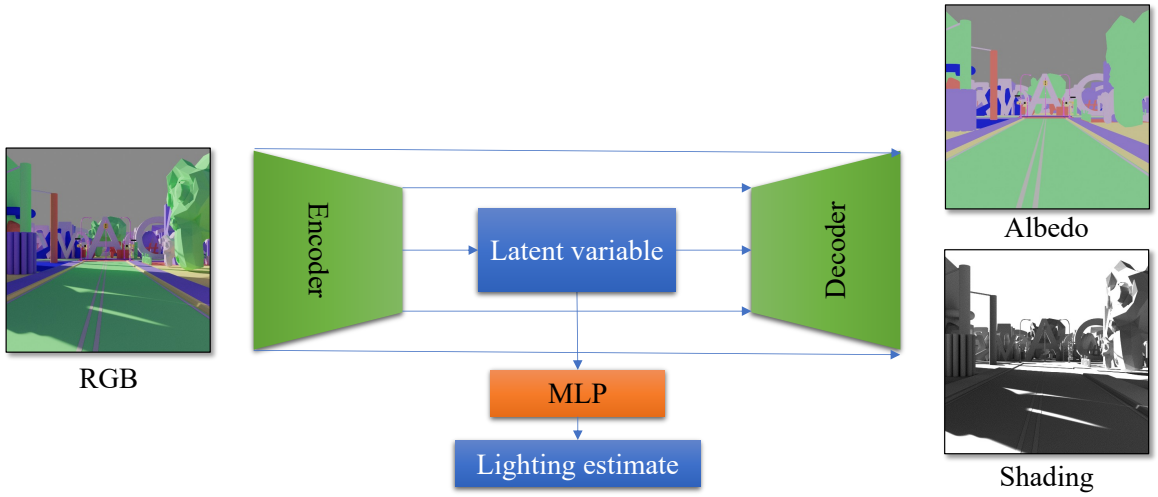


Figure 3.1. Overview of the core of the proposed model: An RGB image is input into the model where the encoder and decoder of the U-Net, denoted in green, estimate the albedo and shading. An additional MLP network, illustrated in orange, located at the bottom, processes the latent variable from the deepest level of the U-Net to estimate the lighting condition of the input image, represented by parameters such as altitude and azimuth.

(a) Reference (b) Same albedo, different shading (c) Different albedo, same shading

Figure 3.2. Triplet Image Examples. Image (b) shares the same albedo as image (a) but differs in lighting conditions, leading to a distinct shading effect. Image (c), while maintaining the same lighting conditions as image (a), features objects with different albedo.

Fig. 3.1 provides an overview of this core structure.

This core network is replicated three times to form a Siamese network. All three networks share the same initial weights and are updated identically during training. The reference image I_1 and the second image I_2 share the same albedo component but differ in shading. Conversely, the third image I_3 shares the same shading component as the reference but has a different albedo component. Fig. 3.2 shows an example of this triplet and Fig. 3.3 illustrates the complete network structure along with examples of inputs and outputs.

Leveraging this relationship among the triplet images, we define the loss function based on the network outputs, A_n, S_n, and $\vec{v}_n$. Initially, the loss function for the intrinsic compo-

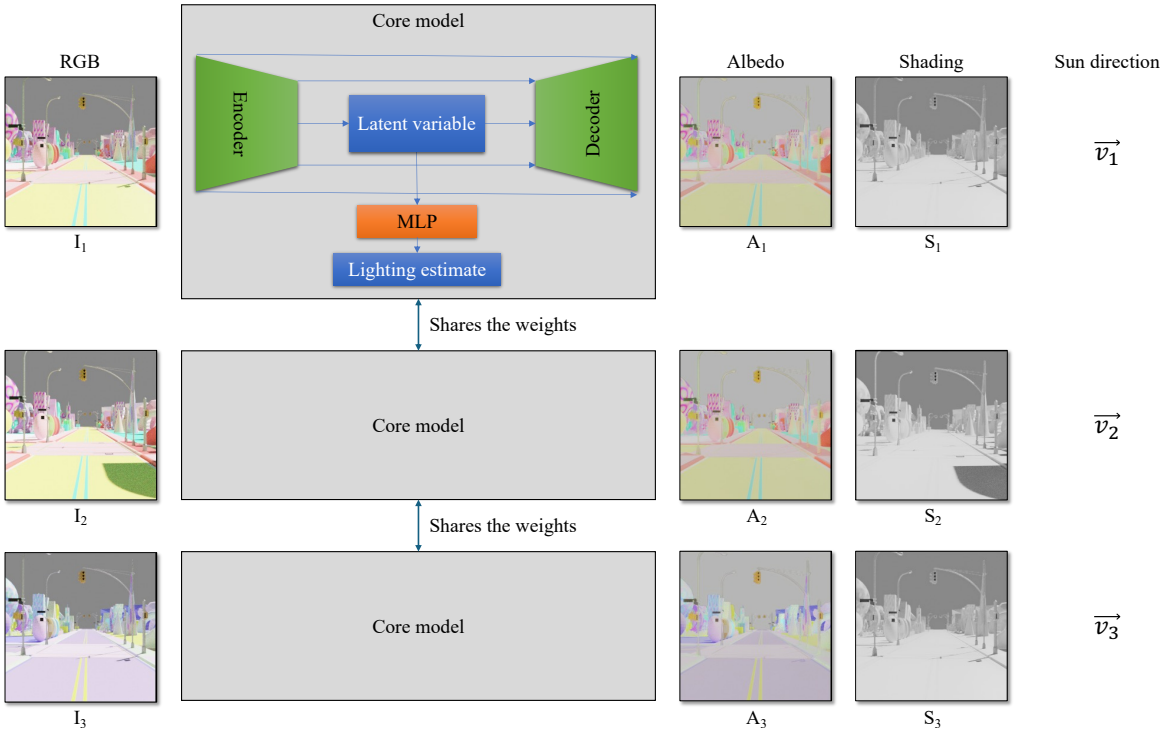


Figure 3.3. Siamese network structure: The core network is replicated three times, sharing the same weights. For a given set of triplet images, the loss is calculated based on their intrinsic decomposition and lighting estimation results.

nents was formulated as follows:

$$\begin{aligned} L_1 &= \text{MSE}(A_2 \odot S_1, I_1) \\ L_2 &= \text{MSE}(A_1 \odot S_2, I_2) \\ L_3 &= \text{MSE}(A_1 \odot S_3, I_1) \\ L_4 &= \text{MSE}(A_3 \odot S_1, I_3) \\ L_{recon} &= L_1 + L_2 + L_3 + L_4, \end{aligned} \tag{3.1}$$

where A_n and S_n indicate the albedo and shading of the n-th image, I_n. We assume the decomposition $I_n = A_n \odot S_n$.

The second component, $L_{lighting}$, is calculated as the cosine dissimilarity between the ground truth lighting direction (v_{gt}) and the predicted lighting direction (v_{pred}), given by:

$$L_{lighting} = 1 - \vec{v}_{gt} \cdot \vec{v}_{pred} / \|\vec{v}_{pred}\|, \tag{3.2}$$

where v_{gt} is a unit vector. Note that the triplet condition does not affect the lighting direction loss, as we assume the availability of the ground truth sun direction for both synthetic and real images.

Therefore, the total loss function, Loss, integrates both reconstruction and lighting accuracy aspects:

$$\text{Loss} = L_{recon} + L_{lighting} \tag{3.3}$$

This loss function framework is crucial for balancing the fidelity of image decomposition against the accuracy of lighting estimation, ensuring a comprehensive approach to the intrinsic image decomposition challenge.

Training of the Siamese network was predominantly conducted on synthetic scenes rendered in Blender, given the challenges in acquiring real-world triplet images.

3.3 Experiment

We designed our experiments to evaluate the effectiveness of our methods. The U-Net architecture utilized for these experiments followed the standard structure, known for its proficiency in image-to-image translation tasks. The network was configured to accept input RGB images at a resolution of 288x288. This design choice was instrumental in maintaining a balance between computational efficiency and the resolution necessary for capturing detailed image features. The output from the U-Net comprised albedo and shading images, matching the resolution of the input images. This ensured a high-fidelity reconstruction of the intrinsic components from the original RGB images.

Complementing the U-Net, a refine network was employed to estimate the lighting conditions. This network was a simple MLP composed of three dense layers. The MLP's design focused on extracting and processing the latent variables obtained from the deepest level of the U-Net, translating them into meaningful lighting parameters such as altitude and azimuth.

For dataset generation, two types of synthetic datasets were created. The first dataset comprised images generated from primitive 3D scenes, designed to test the model's capability in a controlled environment with simpler scene structures. The second dataset involved more complex and realistic 3D scenes, aiming to challenge and evaluate the model's performance under more intricate and varied conditions. Additionally, the method was tested on the KITTI dataset, primarily for lighting estimation. This real-world dataset served as a crucial benchmark, albeit only for the lighting estimation aspect, as it does not provide ground truth data for albedo and shading. The inclusion of the KITTI dataset in the testing phase offered a valuable opportunity to realize the model's limitation in real-world scenarios. Fig. 3.4 is depicting examples of the three datasets.

The network was trained on an Nvidia RTX 2080 for 24 epochs, utilizing an early stopping strategy. The training set included 12,000 triplets, comprising 6,000 from primitive

3D scenes and 6,000 from realistic 3D scenes. For validation, a set of 2,000 triplets was used, equally divided between the two types of 3D scenes. The ground truth sun direction for each image was accurately determined, as all images were rendered from 3D scenes, ensuring reliable and precise data for model evaluation.

3.4 Result

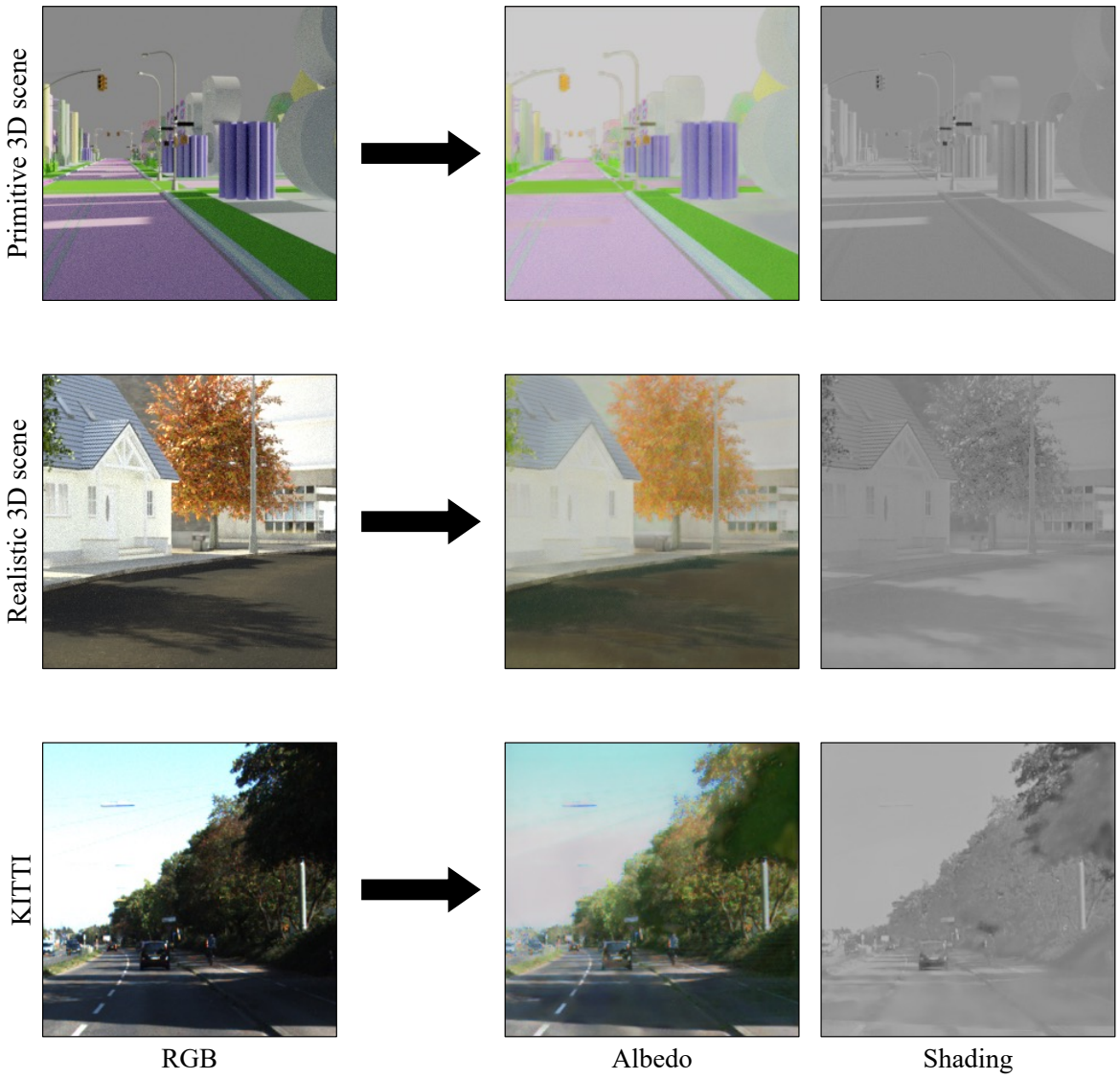


Figure 3.4. Qualitative evaluation of intrinsic image decomposition. The model demonstrates reasonable performance on unseen images from both primitive and realistic synthetic scenes. However, it exhibits limitations when applied to real images, attributed to the significant domain gap.

The experimental phase was structured to evaluate the triplet loss-based Siamese net-

Dataset	Median	Mean	Min	Max
Primitive	12.76	20.07	2.26	99.76
Realistic	54.14	61.69	5.12	147.14
KITTI	51.04	56.11	19.83	131.70

Table 3.1. Summary of angular error statistics in lighting estimation across different test datasets.

work for intrinsic image decomposition and lighting estimation. This section presents the key findings and discusses the implications of the results obtained from the experiments.

Intrinsic Image Decomposition The qualitative assessment of intrinsic image decomposition, as illustrated in Fig. 3.4, reveals a distinct disparity in the performance of the model across different datasets. While the model achieved moderate success in decomposing images from synthetic scenes, both primitive and realistic, it faced considerable challenges with real images from the KITTI dataset. This discrepancy in performance can be largely attributed to the substantial domain gap between synthetic and real-world data. It is important to note the inherent difficulty in obtaining ground truth for albedo and shading in real-world images, a challenge compounded by the triplet-based training scheme used. This scheme, which requires two additional images sharing either albedo or shading with a reference image, further amplifies the complexity of training with real-world data.

Lighting Estimation Regarding lighting estimation, a detailed analysis is presented in Table 3.1, which enumerates the mean, median, minimum, and maximum angular errors across various test datasets. Angular error here refers to the deviation angle between the predicted and the ground truth sun directions, serving as a critical measure of the accuracy in lighting estimation. The error can vary between 0 and 180 degree.

The analysis revealed that the model demonstrated plausible effectiveness primarily on the primitive 3D scene dataset. However, its performance significantly diminished when applied to more complex realistic 3D scenes and the KITTI dataset, indicating a less satisfactory outcome in these contexts. This trend suggests that the latent vector extracted from the deepest level of the U-Net failed to encode critical information pertinent to lighting conditions, challenging a primary hypothesis of the research design. The inability of the latent vector to adequately capture lighting information underscores the necessity for further research. This includes investigating more sophisticated methods for lighting estimation that can more effectively interpret and utilize data from complex real-world scenarios.

These results collectively provide critical insights into the challenges faced by the pro-

posed methods. While the approaches showed potential in certain controlled environments, they revealed significant limitations in more complex and realistic settings. This underscores the need for further research and development in intrinsic image decomposition and lighting estimation, particularly in bridging the gap between synthetic and real-world data, and in developing more robust models that can effectively interpret and utilize deep-layer information for accurate lighting estimation.

3.5 Discussion

The outcomes of this research highlight several crucial insights and limitations. The use of synthetic datasets, while beneficial for controlled conditions, revealed a considerable domain gap in real-world application, particularly evident in the model's performance on the KITTI dataset.

Another notable observation was the limited utility of the latent vector at the deepest level of the U-Net in capturing lighting information. This outcome suggests that the U-Net's skip connections might have played a more substantial role in the decomposition process than the encoding at its deepest layer. An alternate approach, possibly using an autoencoder structure instead of a U-Net, might yield improved results by focusing on the encoding of lighting information without the influence of skip connections.

The recent research by Zhu et al. [51] points to new possibilities in lighting estimation. Future research could take inspiration from such work, exploring the integration of spatially-varying lighting models and enhancing the real-world applicability of these methods. Additionally, the development of more sophisticated network structures and the incorporation of physically-based constraints could further improve the accuracy and applicability of lighting estimation in complex scenarios.

Chapter 4

Spatio-Temporal Outdoor Lighting Aggregation on Image Sequences

The material presented in this chapter is derived from published works. In addition to the introduction presented here, the comprehensive material from [7] and its expanded version [52] are incorporated into other sections of this chapter, as well as the abstract, introduction (chapter 1), and conclusion (chapter 5) of this thesis.

Deep learning has dramatically transformed the landscape of computer vision, enabling the extraction and interpretation of intricate information from visual data. Among its numerous applications, one significant area is the estimation of lighting conditions from imagery, particularly in outdoor environments [19, 49, 53–57]. This task is pivotal for a range of applications, most notably in the realm of augmented reality (AR) and holistic scene understanding. In AR, for instance, the accurate and realistic rendering of virtual objects into real-world images necessitates a precise understanding of the prevailing lighting conditions [58]. However, the inherent complexity of outdoor scenes, coupled with the ill-posed nature of the problem, presents significant challenges. The interplay of material properties, geometric configurations, and varying lighting conditions can lead to identical pixel representations, making the task of accurately estimating lighting from a single image particularly challenging without imposing additional constraints.

The importance of accurate lighting estimation extends beyond AR. It plays a crucial role in enhancing the realism of virtual scenes in various domains, including cinematography, architectural visualization, and video games. Furthermore, in the field of computer vision, accurate lighting information is instrumental for tasks such as depth-from-mono estimation [59–61], where convolutional neural networks have shown improved performance when provided with realistic shadow information. Despite these applications, the estimation of lighting conditions has remained a challenging area of research due to the complexities involved in deciphering the myriad of lighting cues present in natural scenes.

Our research endeavors to address these challenges by focusing on the estimation of environment lighting, which is a foundational step towards comprehensive scene understanding. The significance of environment lighting estimation lies in its ability to provide a cohesive understanding of the lighting dynamics in outdoor scenes, which are crucial for applications like AR, where the perception of depth, material, and spatial relationships heavily depends on accurately rendered lighting and shadows [51, 58, 62–64]. Traditional methods in this domain have predominantly concentrated on either reconstructing sky map textures, identifying sun positions from RGB images, or deducing multiple light source locations using material information. These methods, however, face limitations due to their susceptibility to noise and the prevalence of outliers in the data obtained from individual images.

In our first study [7] (hereafter referred to as the *'Four-Stage Approach'*), we introduced a pioneering approach for robustly estimating the global sun direction in outdoor environments by exploiting the spatial and temporal coherency present in lighting conditions. This approach involves a new four-stage post-processing method that combines spatial and temporal filtering with outlier detection. By analyzing sub-views of an image sequence, our method effectively samples across both angular and temporal domains. This dual-domain sampling offers a twofold advantage: it filters out noise and detects outliers, and it enables the neural network-based lighting estimator to become invariant to various imaging parameters, such as size, aspect ratio, and camera focal length. The key contributions of this study are a single image-based sunlight estimation using a deep artificial neural network with a convolutional block attention module, and a unique tunable statistical post-processing approach, which together mark a significant step forward in outdoor lighting estimation.

Building on this foundation, our second study [52] (hereafter referred to as the *'End-to-End Approach'*) further advances the field by introducing an innovative end-to-end model that supersedes the initial statistical post-processing method with a Transformer architecture. This new model streamlines the estimation process by eliminating the necessity for intricate hyperparameter tuning associated with previous methods. Additionally, it incorporates a new handcrafted positional encoding mechanism, designed to effectively encode local and global camera angles for spatio-temporal aggregation. Our extension not only enhances the efficiency of the estimation process but also significantly improves the realism of the lighting model. By adopting the Lalonde-Matthews outdoor illumination model [19], our method provides a more comprehensive and realistic estimation of lighting conditions. The contributions from this study include the application of an attention-based model for the task of lighting estimation, a pioneering positional encoding method tailored for spatio-temporal aggregation, and a performance that surpasses existing state-of-the-art methods.

In conclusion, our journey through the complex landscape of spatio-temporal outdoor lighting estimation represents a significant advancement in the field. We have transitioned

from a single image-based approach with post-processing to a sophisticated end-to-end model, demonstrating the profound potential of deep learning in addressing intricate, ill-posed problems. These contributions not only solidify the importance of precise lighting estimation in various applications but also pave the way for more realistic augmented reality experiences and a deeper understanding of outdoor scenes.

4.1 Related Work

Estimation of outdoor lighting conditions has been extensively studied due to its importance in computer graphics and computer vision applications [65, 66]. Related techniques can be categorized into two parts, one that analyzes a single image [49, 53–55, 57, 67, 68] and the other that utilizes a sequence of images [19, 62, 64, 69].

4.1.1 Single Image

Hold-Geoffroy et al. [53] proposed a method that estimates outdoor illumination from a single low dynamic range image using a convolutional neural network [38] (CNN). The network was able to classify the sun location on 160 evenly distributed positions on the hemisphere and estimated other parameters such as sky turbidity, exposure, and camera parameters.

Analyzing outdoor lighting conditions is further developed in [56] where they incorporated a more delicate illumination model [19]. The predicted parameters were evaluated numerically with the ground truth values and rather qualitatively assessed by using the render loss.

Jin et al. [55] and Zhang et al. [57] also proposed single image based lighting estimation methods. While their predecessors [53, 56] generated a probability distribution of the sun position on the discretized hemisphere, the sun position parameters were directly regressed from their networks. Recently, Zhu et al. [51] combined lighting estimation with intrinsic image decomposition. Although they achieved a noticeable result in sun position estimation on synthetic datasets, we could not compare them to ours because their method utilizes intrinsic images which are unavailable for real scene videos.

4.1.2 Multiple Images

The above lighting estimation methods based on a single image often suffer from insufficient cues to determine a lighting condition, such as when a given image is completely shadowed. Therefore, several attempts were made to increase the accuracy and robustness by taking the temporal domain into account [19, 62, 64].

For example, in the outdoor illumination estimation method presented by Madsen et al. [70], the authors estimated the trajectory of the sun and its variable intensity from a sequence of images. Under the assumption that a static 3D model of the scene is available, they designed a rendering equation-based [8] optimization problem to determine the continuous change of the lighting parameters. The method introduced in [69] extracts a set of features from each image frame and uses it to estimate the relative changes of the lighting parameters in an image sequence. Their method can handle moving cameras and generate time-coherent augmentations. However, the estimation process utilized only two consecutive frames and assumed that the sun position is given in the form of GPS coordinates and timestamps [71].

The lighting condition estimation is also crucial in augmented reality where virtual objects are realistic when they are rendered in the background image using the correct lighting conditions. Lu et al. [66], for instance, estimated a directional light vector from shadow regions and the corresponding objects in the scene to achieve realistic occlusion with augmented objects.The estimation performance depends solely on the segmentation of the shadow region and the finding of related items. Therefore, the method may be challenging if a shadow-casting object is not visible in the image. Madsen and Lal [64] utilize a stereo camera to extend [70] further. They estimated sky and sun variations over an image sequence using the sun direction calculated from the GPS coordinates and time stamps. The estimates are then combined with shadow detection algorithms to generate plausible augmented scenes with appropriate shading and shadows.

Recently, several attempts have been made to use auxiliary information to estimate lighting conditions [63, 72]. Such information may result in better performance but only with a trade-off in generality. Kán and Kaufmann [63] proposed a single RGB-D image-based lighting estimation method for augmented reality applications. They used synthetically generated scenes to train a deep neural network that maps the angular coordinates of the main light source in the scene. Outlier removal and temporal smoothing processes were applied to achieve temporal consistency of the method. However, this method was demonstrated only on static-view images. Our method, on the other hand, improves its estimates by aggregating observations from different points of view. We illustrate the consistency gained from our new design by augmenting virtual objects in consecutive frames.

4.2 Method

In this chapter, we detail the methodologies of two advanced spatio-temporal lighting estimation techniques. The first approach, known as the ‘Four-Stage Approach’, estimates lighting conditions on single images, followed by a two-part post-processing stage. Ini-

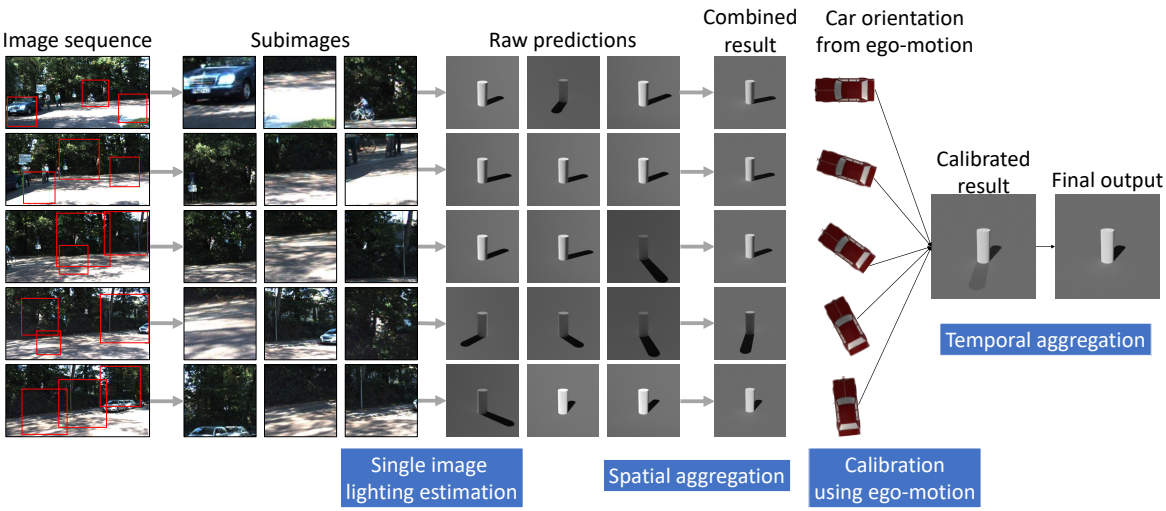


Figure 4.1. Spatio-temporal outdoor lighting aggregation on an image sequence: individual estimates from each generated subimage are combined in the spatial aggregation step. Spatial aggregation results for each image in the sequence are then calibrated using camera ego-motion data and further refined in the temporal aggregation step to generate the final lighting estimate for the sequence.

tially, noisy lighting estimates are aggregated spatially, then, through a calibration step, we adjust for the angle of the ego vehicle, and finally, we apply post-processing in the temporal domain. This method efficiently aggregates spatially processed observations over time, enhancing the robustness of lighting estimation.

Building upon this, our second approach, the 'End-to-End Approach', streamlines the process by eliminating the separate post-processing stage. Instead, it integrates a Transformer network that employs a spatio-temporal attention mechanism, providing a more cohesive and efficient means of aggregating lighting information. This approach also incorporates the advanced Lalonde-Matthews sun-sky lighting model, offering a superior and more realistic estimation of outdoor lighting conditions. The evolution from the Four-Stage to the End-to-End Approach represents a significant advancement in spatio-temporal lighting estimation, shifting from a process with separate stages to a more integrated and sophisticated model.

4.2.1 Four-Stage Approach

We take advantage of different aspects of previous work and refine them into our integrated model. As illustrated in Fig. 4.2, our model is composed of four subprocesses. We first randomly generate several small subimages from an input image and upsample them to a fixed size. Since modern cameras are capable of capturing fine details of a scene, we found

that lighting condition estimation can be done on a small part of an image. These spatial samples obtained from one image all share the same lighting condition and therefore yield more robustness compared to a single image view. Then, we train our lighting estimation network on each sample to obtain the global lighting for a given input image.

After the network estimates the lighting conditions for the spatial samples, we perform a spatial aggregation step to get a stable prediction for each image. Note that the estimate for each frame is based on its own camera coordinate system. Our third step is to unify the individual predictions into one global coordinate system using the camera ego-motion. Lastly, the calibrated estimates are combined in the temporal aggregation step. The assumption behind our approach is that distant sun-environment lighting is invariant to the location the picture was taken and that the variation in lighting direction is negligible for short videos. Through the following sections, we introduce the details of each submodule.

Lighting Estimation

There have been several sun and sky models to parameterize outdoor lighting conditions [19, 73]. Although those methods are potentially useful to estimate complex lighting models consistently, in this work we focus only on the most critical lighting parameter: the sun direction. The rationale behind this is that ground-truth training data can easily be generated for video sequences with GPS and timestamp information (e.g., KITTI dataset [2]). Therefore, the lighting estimation network's output is a 3D unit vector $\vec{v}_{pred}$ pointing to the sun's location in the camera coordinate system.

Unlike our predecessors [53, 56], we design our network as a direct regression model to overcome the need for a sensitive discretization of the hemisphere. The recent work of Jin et al. [55] presented a regression network estimating the sun direction in spherical coordinates (altitude and azimuth). Our method, however, estimates the lighting direction using Cartesian coordinates and does not suffer from singularities in the spherical parametrization and the ambiguity that comes from the cyclic nature of the spherical coordinates.

Since we train our network in a supervised manner, the loss function is defined to compare the estimated sun direction with the ground truth $\vec{v}_{gt}$:

$$L_{cosine} = 1 - \vec{v}_{gt} \cdot \vec{v}_{pred} / ||\vec{v}_{pred}||, \tag{4.1}$$

with the two adjacent unit vectors having their inner product close to 1. To avoid the uncertainty that comes from the vectors pointing the same direction with different lengths, we apply another constraint to the loss function:

$$L_{norm} = (1 - ||\vec{v}_{pred}||)^2. \tag{4.2}$$

The last term of the loss function ensures that the estimated sun direction resides in the upper hemisphere because we assume the sun is the primary light source in the given scene:

$$L_{hemi} = max(0, -z_{pred}), \tag{4.3}$$

where z_{pred} is the third component of $\vec{v}_{pred}$, indicating the altitude of the sun. The final loss function is simply the sum of all terms as they share a similar range of values:

$$L_{sun} = L_{cosine} + L_{norm} + L_{hemi}. \tag{4.4}$$

Spatial Aggregation

Using our lighting estimator, we gather several lighting condition estimates from different regions of the image. Some of those estimates may contain larger errors due to insufficient information in the given region to predict the lighting condition. We refer to such estimates as outliers. Our method's virtue is to exclude anomalies that commonly occur in single image-based lighting estimation techniques and deduce the best matching model that can explain the inliers.

Among various outlier removal algorithms, we employ the isolation forest (iForest) algorithm [74]. The technique is specifically optimized to isolate anomalies instead of building a model of inliers and eliminate samples not complying with it. In essence, the iForest algorithm recursively and randomly splits the feature space into binary decision trees (hence forming a forest). Since the outliers are outside of a potential inlier cluster, a sample is classified as an outlier if the sample's average path length is shorter than a threshold (*contamination ratio* [75]). We determine this value empirically and use it throughout all results.

On the remaining inliers, we apply the *mean shift algorithm* [76] to conjecture the most feasible lighting parameters. Unlike naive averaging over all inliers, this process further refines the lighting estimate by iteratively climbing to the maximum density in the distribution. Another experimentally discoverable parameter *bandwidth* determines the size of the Gaussian kernel to measure the samples' local density gradient. In the proposed method, we set the bandwidth as the median of all samples' pairwise distances. By moving the data points iteratively towards the closest peak in the density distribution, the algorithm locates the highest density within a cluster, our spatial aggregation result. We compare various aggregation methods in the ablation study in 4.4.1.

Calibration

Since our primary goal is to assess the sun direction for an input video, we perform a calibration step to align the estimates because the sun direction determined from each image in a sequence is in its own local camera coordinate system. The camera ego-motion data is

necessary to transform the estimated sun direction vectors into the world coordinate system. We assume the noise and drift in the ego-motion estimation is small relative to the lighting estimation. Hence, we employ a state-of-the-art structure-from-motion (SfM) technique such as [34] to estimate the ego-motion from an image sequence. Then there exists a camera rotation matrix R_f for each frame f and the resulting calibrated vector $\hat{\vec{v}}_{pred}$ is computed as $R_f^{-1} \cdot \vec{v}_{pred}$.

Temporal Aggregation

Having the temporal estimates aligned in the same global coordinate system, we consider them as independent observations of the same lighting condition in the temporal domain. Although the lighting estimates from our regression network are not necessarily independent for consecutive video frames, natural image sequences, as shown empirically in our experiments, reveal a large degree of independent noise in the regression results, which is however polluted with a non-neglectable amount of outliers. Consequently, we apply a similar aggregation strategy as in the spatial domain also for the temporal domain. Therefore, the final output of our pipeline, the lighting condition for the given image sequence, is the mean shift algorithm's result on the inliers from all frames of the entire image sequence.

4.2.2 End-to-End Approach

In developing our advanced model, we have integrated key elements from prior research, culminating in a system that synergizes two distinct networks: a ResNet18 [6] and a Transformer network [42], as shown in Fig. 4.2. This model begins by extracting numerous small subimages from an image sequence, utilizing the high-resolution capabilities of contemporary cameras to focus on minute scene details for lighting estimation. The process of analyzing these smaller sections across different sequences allows us to capture a wide range of observations, all contributing to a comprehensive understanding of the global lighting conditions. This methodology is rooted in empirical findings, demonstrating the effectiveness of assessing lighting from multiple small segments within an image sequence.

All image crops are passed through the backbone network and projected to a sequence of patch embeddings. We then add an orientation-invariant positional encoding and pass the sequence to our transformer network. Through the attention layers, the noisy spatio-temporal observations can be effectively aggregated to a final estimate. Weighted features are delivered to a dense layer that produces the estimated *Lalonde-Matthews* illumination model parameters. The sun direction estimates are formulated in their own camera coordinate systems. We compensate the camera yaw angle of each subimage in order to obtain aligned estimates in a unified global coordinate system. Our final prediction is given as the

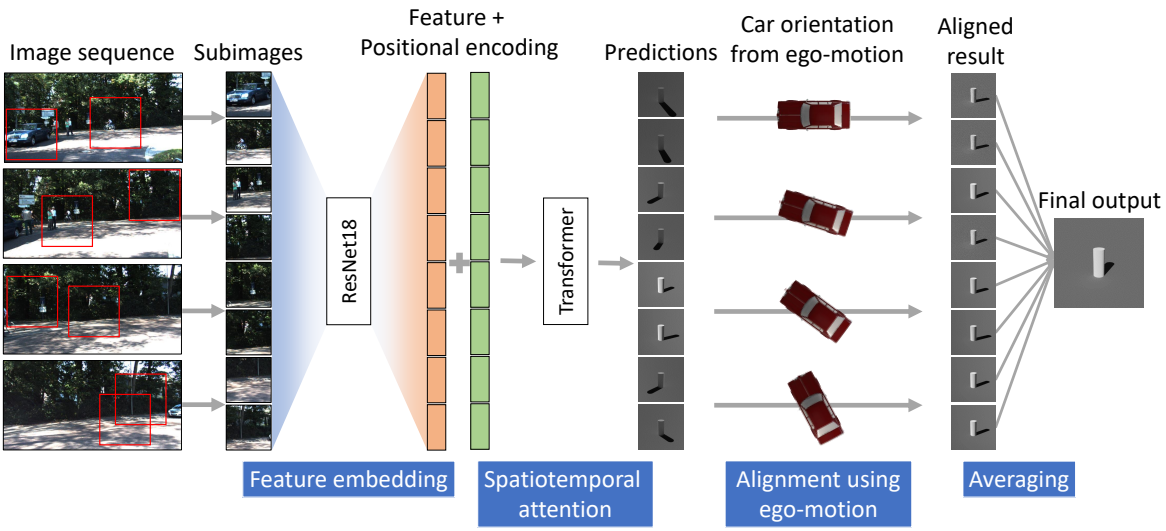


Figure 4.2. Spatio-temporal outdoor lighting aggregation on an image sequence: feature vectors are extracted from subimages using a ResNet18 network structure. Using an absolute positional encoding, our transformer network performs spatio-temporal attention. Individual estimates made in each camera coordinate system are aligned using camera yaw angle data and fused to yield the lighting estimation for the sequence.

average of all estimates. Note that the sky parameters of the *Lalonde-Matthews* model do not require the alignment step, as they do not vary with respect to the camera yaw angle. The assumption behind our spatio-temporal aggregation is that distant sun-environment lighting can be considered invariant for small-scale translations (e.g., driving) and that the variation in lighting direction is negligible for short videos. Through the following sections, we introduce the details of our method.

Lighting Estimation

In our research, we have employed sophisticated sun and sky models to parameterize outdoor lighting conditions, notably the *Hosek-Wilkie* sky model [73] and the *Lalonde-Matthews* outdoor illumination model [19]. Advancing from our previous methods, this work uniquely focuses on predicting the parameters of the *Lalonde-Matthews* model. This model, denoted as f_{LM}, intricately captures the luminance of outdoor illumination as a function of light direction $\mathbf{l}$, comprising both sun (f_{sun}) and sky (f_{sky}) components. It operates on a compre-

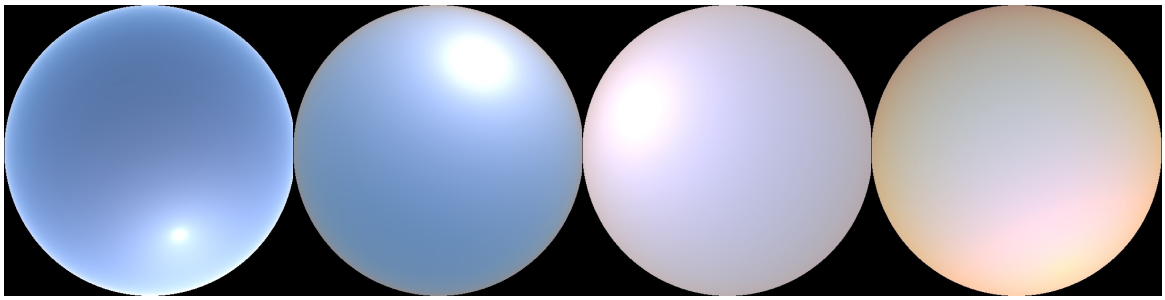


Figure 4.3. Sky textures generated by the *Lalonde-Matthews* model with various sets of 11 parameters, each depicting a sky hemisphere where the center of the circle corresponds to the zenith.

hensive set of 11 parameters:

$$
\begin{aligned}
f_{LM}(\mathbf{l}; q_{LM}) &= \mathbf{w}_{sun} f_{sun}(\mathbf{l}; \beta, \kappa, \mathbf{l}_{sun}) + \mathbf{w}_{sky} f_{sky}(\mathbf{l}; t, \mathbf{l}_{sun}), \\
f_{sun}(\mathbf{l}; \beta, \kappa, \mathbf{l}_{sun}) &= \exp(-\beta \exp(-\kappa / \cos \gamma_{\mathbf{l}})), \\
f_{sky}(\mathbf{l}; t, \mathbf{l}_{sun}) &= f_P(\theta_{\mathbf{l}}, \gamma_{\mathbf{l}}, t), \\
q_{LM} &= \{\mathbf{w}_{sun}, \mathbf{w}_{sky}, \beta, \kappa, t, \mathbf{l}_{sun}\}
\end{aligned}
$$

These parameters include $\mathbf{w}_{sun} \in \mathbb{R}^3$ and $\mathbf{w}_{sky} \in \mathbb{R}^3$ representing mean sun and sky colors, sun shape descriptors (β, κ), sky turbidity t, and the sun's position $\mathbf{l}_{sun} = [\theta_{sun}, \phi_{sun}]$. The angle $\theta_{\mathbf{l}}$ and $\gamma_{\mathbf{l}}$ measure the zenith angle of the light direction $\mathbf{l}$ and the orientation of $\mathbf{l}$ relative to the sun position $\mathbf{l}_{sun}$ respectively, and f_P is based on the Preetham sky model [77]. For more details, please refer to Appendix A. Fig. 4.3 displays sky textures generated by the *Lalonde-Matthews* model.

To accurately estimate the sun direction, a critical parameter in this model, we adopted a direct regression approach using Cartesian coordinates for sun direction estimation. This method, distinct from previous models [55, 57] that relied on spherical coordinates, avoids the complications of singularities and cyclic ambiguities inherent in spherical parametrization.

For the sun direction estimation within the *Lalonde-Matthews* model, we utilize the same loss function, L_{sun} in Eq. 4.4, as employed in the Four-Stage Approach. This ensures consistency in our methodology while focusing our advancements on other aspects of the model, particularly the integration with the *Lalonde-Matthews* illumination parameters.

For the remaining parameters, we apply the mean squared error (MSE) to the predicted

values and the normalized ground truth values as in [55]:

$$L_{\mathbf{w}_{sun}} = \frac{1}{3}\left\|\mathbf{w}_{sun}^{pred} - \mathbf{w}_{sun}^{gt}\right\|_2^2 \tag{4.5}$$

$$L_{\mathbf{w}_{sky}} = \frac{1}{3}\left\|\mathbf{w}_{sky}^{pred} - \mathbf{w}_{sky}^{gt}\right\|_2^2 \tag{4.6}$$

$$L_{beta} = \left\|\beta^{pred} - \beta^{gt}\right\|_2^2 \tag{4.7}$$

$$L_{kappa} = \left\|\kappa^{pred} - \kappa^{gt}\right\|_2^2 \tag{4.8}$$

$$L_{t} = \left\|t^{pred} - t^{gt}\right\|_2^2 \tag{4.9}$$

$$L_{param} = \frac{1}{5}\Big[L_{\mathbf{w}_{sun}} + L_{\mathbf{w}_{sky}} + L_{beta} + L_{kappa} + L_{t}\Big] \tag{4.10}$$

Since the two loss functions L_{sun} and L_{param} have similar magnitudes, we define the final loss function as the sum of them:

$$L_{light} = L_{sun} + L_{param}. \tag{4.11}$$

Attention-based Aggregation

In order to extract robust estimates from noisy observations, the aggregation process described in [7] relies heavily on statistical filtering utilizing an outlier removal combined with the meanshift algorithm. However, this approach requires manual hyperparameter tuning with handcrafted selection criteria. We extend this work by replacing the aggregation step with a purely end-to-end attention driven pipeline. The overview of our approach is illustrated in Figure 4.2.

We take inspiration from [43] for our network design and adopt their hybrid architecture for our task. This includes self attention using multi-head attention layers [42] and preprocessing images with a convolutional neural network. Given a temporal sequence of k images, we first select n spatially randomized crops for each frame as done in our previous work [7]. For each crop, we apply a ResNet18 [6] encoder to extract feature embeddings. Each embedded patch is fed as input to our transformer module for aggregation. The virtue of the transformer network is that it can associate observations from different space and time given a proper positional encoding. Since all image patches share the same lighting condition and we assume we know their relative orientation due to the ego-motion estimation, the Transformer's attention mechanism inherently learns to filter the noisy patch-wise predictions. However, we need to provide the relative orientation of the patches in order to make the light estimation invariant to the camera orientation, which we achieve via the *positional encoding*.

Orientation-invariant Positional Encoding

Solely relying on image features enables only to estimate the lighting in the local camera frame. However, we need to fuse the estimates into a global reference frame in order to relate different subimages. Since we assume sun-lighting, only the directional component of a recorded camera image is relevant to align different frames. We inject this camera orientation into the image features via a positional encoding. However, we encode only the yaw angle of the camera rotations (the rotation around the ground-plane surface normal) since pitch and roll angles are naturally captured in the image features of outdoor images (e.g., horizon). Further, we also encode the 2D position of the subimages cropped from the source frame independent of the intrinsic camera projection, i.e., in terms of viewing angles ϕ in the corresponding horizontal and vertical field of views. For example, the top left pixel gets a coordinate of $\left(-\frac{\sphericalangle_h}{2},\ \frac{\sphericalangle_v}{2}\right)$ for a pinhole camera model with a field of view of $\sphericalangle_h$ and $\sphericalangle_v$ horizontally and vertically, respectively. To this end, we concatenate the *2D angular image coordinate* and the (temporal) *camera rotation angle* and apply our positional encoding. We use an absolute positional encoding, i.e.

$$x_i^{enc} \longleftarrow x_i + p_i \ , \tag{4.12}$$

where the positional encoding p_i and the subimage feature vector $x_i \in \mathbb{R}_x^d$ are superimposed. Similar to [42] we use a fixed encoding of sine and cosine functions with different frequencies.

Since our positional encoding scheme encodes angles, it has to fulfill the following two conditions: 1) *periodicity* - the transition from the encoding of $359°$ to the encoding of $0°$ should be as smooth as the transition from $0°$ to $1°$ and 2) *uniqueness* - each angle should have a unique encoding. We present our cyclic positional encoding, satisfying those conditions, by using nested trigonometic functions as follows:

$$\begin{aligned} PE\left(\phi,\ 2i\right) &= \sin\left(\sin\left(\phi\right) \cdot \alpha / 10000^{2i/d}\right) \\ PE\left(\phi,\ 2i+1\right) &= \sin\left(\cos\left(\phi\right) \cdot \alpha / 10000^{2i/d}\right) , \end{aligned} \tag{4.13}$$

where $i \in [0,\ \frac{d}{2})$ and d denotes the depth of the positional encoding. Note that α is an empirically determined parameter, which controls the width of the nonzero area of the encoding. The periodicity comes from the nested trigonometric function, whereas the uniqueness is established by interlacing the two functions. Fig. 4.4 shows the positional encoding generated by the above function.

The resulting positional encoding of a subimage is the stacked vector of the three cyclic positional encodings. Note that the depth parameter d is carefully determined so that the depth of the stacked vector matches the channel size of the transformer network.

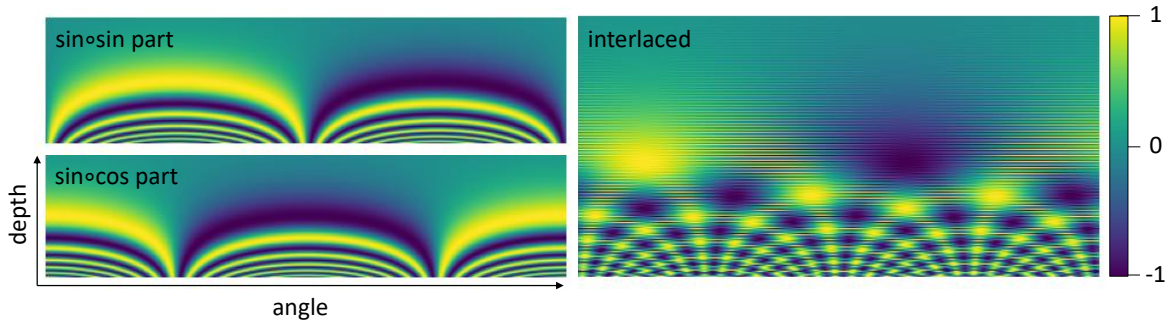


Figure 4.4. Cyclic positional encoding for angle $\phi \in [0,\ 2\pi]$. The periodicity of our encoding scheme is clearly visible on the left side images while their interlaced result on the right side shows its uniqueness for each angle.

Alignment

In the End-to-End Approach, the alignment of the sun direction component within the 11-dimensional lighting parameter vector is crucial, akin to the calibration step in the Four-Stage Approach. This is because these estimates are initially made in each image's local camera coordinate system, but for a comprehensive analysis, we need them represented in the global coordinate system.

To align these sun direction estimates, we employ the same structure-from-motion (SfM) technique, as used in the Four-Stage Approach, referenced in [34]. This technique is instrumental in estimating the ego-motion of the image sequence, thereby transforming the local sun direction vectors into the global coordinate system.

For each frame, identified as f, we calculate the camera rotation matrix R_f. We then align the sun direction vectors using the calculation $\hat{\vec{v}}_{pred} = R_f^{-1} \cdot \vec{v}_{pred}$. This alignment is essential to ensure that the sun direction estimates from each frame are accurately oriented within the global coordinate system.

The key distinction in our End-to-End Approach lies in the final step of our process. Here, we take the mean of these aligned estimates across all 11 parameters to formulate our final prediction. This averaging step is applied not just to the sun direction estimates but to the entire set of lighting parameters, reflecting the holistic and integrated nature of the End-to-End Approach.

4.3 Experiments

4.3.1 Four-Stage Approach

Datasets

One of the common datasets considered in the outdoor lighting estimation methods is the SUN360 dataset [3]. Several previous methods utilized it in its original panorama form or as subimages by generating synthetic perspective images [53]. We follow the latter approach since we train our network using square images. We first divide 20 267 panorama images into the training, validation, and test sets with a 10:1:1 ratio. For the training and the validation sets, 8 subimages from each panorama are taken by evenly dividing the azimuth range. To increase the diversity, 64 subimages with random azimuth values are generated from each panorama in the test set. Note that we introduce small random offsets on the camera elevation with respect to the horizon in $[-10^\circ, 10^\circ]$ and randomly select a camera field of view within a range $[50^\circ, 80^\circ]$. The generated images are resized to 256×256. In this way, we produced 135 128, 13 504, and 108 032 subimages from 16 891, 1688, and 1688 panoramas for the training, validation, and test sets, respectively. The ground truth labeling was given by the authors of [56].

The well-known KITTI dataset [2] has also attracted our attention. Since the dataset is composed of several rectified driving image sequences and provides the information required for calculating the ground truth sun directions [71], we utilize it for both training and test. Specifically, since the raw data was recorded at $10\,Hz$, we collect every 10^{th} image to avoid severe repetition and split off five randomly chosen driving scenes for validation and test set. The resulting training set is composed of 3630 images. If we train our network using only one crop for each KITTI image, the network is likely to be biased to the SUN360 dataset due to the heavy imbalance in the amount of data. To match the number of samples, we crop 32 subimages from one image by varying the cropping location and the crop size. Each image in the test set is again cropped into 64 subimages and the cropped images are also resized to 256×256. In total, we train our network on about 250 000 images. The exact numbers of samples are presented in Table 4.1 and Fig. 4.5 illustrates examples of the two datasets.

Implementation Details

Our lighting estimation model is a regression network with convolution layers. It accepts an RGB image of size 256×256 and outputs the sun direction estimate. We borrow the core structure from ResNeXt [4] and carefully determine the number of blocks, groups, and filters as well as the sizes of filters under extensive experiments. As illustrated in Fig. 4.6,

Dataset		SUN360	KITTI
Training	Data	16 891	3630
	Subimg	135 128	116 160
Test	Data	1688	281
	Subimg	108 032	17 984

Table 4.1. Number of data and subimages for training and test

the model is roughly composed of 8 bottleneck blocks, each of which is followed by a convolutional block attention module [5]. In this way, our network is capable of focusing on important spatial and channel features while acquiring resilience from vanishing or exploding gradients by using the shortcut connections. A global average pooling layer is adopted to connect the convolution network and the output layer and serves as a tool to mitigate possible overfitting [78]. The dense layer at the end then refines the encoded values into the sun direction estimate.

We train our model and test its performance on the SUN360 and the KITTI datasets (see Table 4.1). In detail, we empirically trained our lighting estimation network for 18 epochs using early stopping. The training was initiated with the Adam optimizer [79] using a learning rate of 1×10^{-4} and the batch size was 64. It took 12 hours on a single Nvidia RTX 2080 Ti GPU. Prediction on a single image takes 42 ms. Our single image lighting estimation and spatial aggregation modules are examined upon 108 032 unobserved SUN360 crops generated from 1688 panoramas. The whole pipeline including the calibration and temporal aggregation modules is analyzed on five unseen KITTI sequences composed of 281 images.

Figure 4.5. Examples of the two datasets [2, 3]. From the original image (*top*), we generate random subimages (*bottom*).

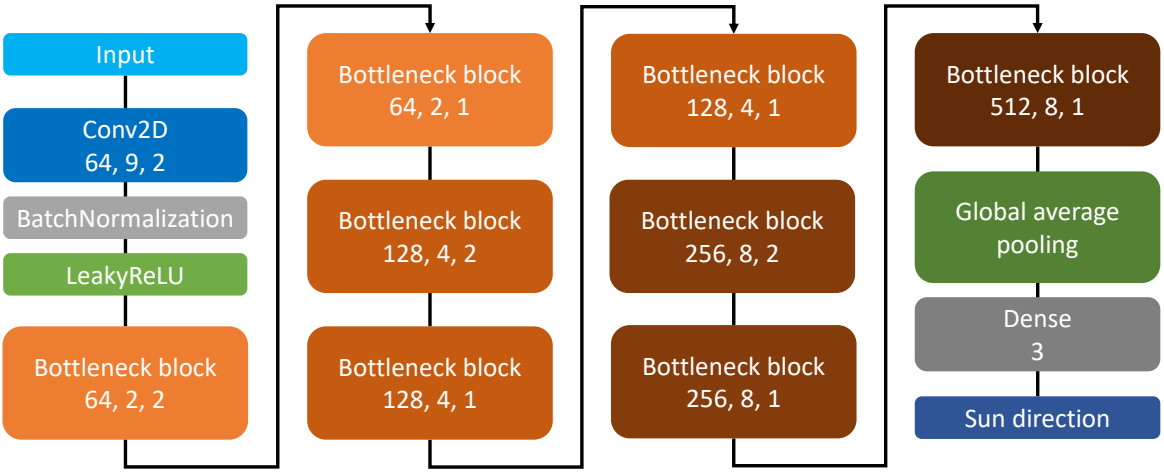


Figure 4.6. The proposed lighting estimation network. The numbers on the *Conv2D* layer indicate the number of filters, the filter size, and the stride, whereas the numbers on each *Bottleneck block* depict the number of 3×3 filters, the cardinality, and the stride. A *Bottleneck block* is implemented following the structure proposed in [4] except for a convolutional block attention module [5] attached at the end of each block.

4.3.2 End-to-End Approach

Datasets

We choose two datasets for evaluation: KITTI [2] and SUN360 [3]. KITTI is a popular dataset for autonomous driving. It consists of multiple driving sequences with rectified images and has additional annotations for determining the ground-truth sun directions [71]. This makes it an ideal candidate to test our method on everyday driving scenes. For our experiments we create a random *train-val-test* split composed of 47-5-5 driving scenes. This results in 33 889, 3508, and 3457 images, respectively. Note that this *scene* is different from the *sequence* we give to the network. Since we generate a *sequence* by randomly selecting eight frames from the same *scene* during the training and inference, there are 4208, 427, and 432 sequences for the *train-val-test* split, respectively. (see Table 4.2). For the sampling in the spatial domain, four subimages are randomly cropped from each frame image while allowing overlapping. Our pipeline estimates the global sun direction from this spatio-temporal sequence of 32 images. Since KITTI does not provide ground truth Lalonde-Matthews lighting model parameters, we omit the loss for other lighting parameters (L_{param}). Therefore, the loss function becomes $L_{light} = L_{sun}$.

The SUN360 dataset is another common dataset considered for outdoor lighting estimation methods because 1) it provides diverse environments and 2) there is a labeling of the parameters of the Lalonde-Matthews lighting model [56]. Several previous methods

Dataset	Training		Validation		Test	
	Sequences	Images	Sequences	Images	Sequences	Images
SUN360	10000	160000	1000	16000	1000	16000
KITTI	4208	33889	427	3508	432	3457

Table 4.2. Number of data in our datasets

used it in its original panorama form or as subimages by generating synthetic perspective images [53]. We followed the latter approach, which has also been used in our preliminary work [7] where we examined the performance improvement arising from spatial aggregation.

In this paper, we propose to build an artificial image sequence from a panorama so that we can examine and compare our method's performance with previous works. Specifically, we simulate a camera motion without translation by generating a set of synthetic perspective images with a fixed field of view and randomized camera yaw and pitch angles. By doing so, we can perform the spatio-temporal aggregation on the SUN360 dataset in the same manner as on KITTI. We start with dividing 12 000 panorama images into the training, validation, and test sets with a 10:1:1 ratio. From each panorama, a sequence of 16 perspective images with random yaw angles is generated while allowing overlapping. We want to have the data from both datasets as similar as possible. Therefore, we match the horizontal and vertical field of views and set the numbers of random frames and subframes to 8 and 4 respectively. Since there are 16 frames for each panorama, a sequence of 8 frames has C_8^{16} different combinations, resulting in great diversity. Note that we also introduce small random offsets on the camera elevation with respect to the horizon in $[-10^\circ, 10^\circ]$. The generated images are resized to 1220×370 to match the size of the KITTI images. In this way, we produced 160 000, 16 000, and 16 000 images from 10 000, 1000, and 1000 panoramas for training, validation, and test sets, respectively. The exact numbers of panoramas and images are presented in Table 4.2, and Fig. 4.5 illustrates examples from the two datasets.

Implementation Details

As illustrated in Fig. 4.7, our lighting estimation model consists of a ResNet18 network and a transformer network, followed by dense layers converting a feature vector of dimension 512 to the estimates for the 3D sun direction and other lighting parameters (only applicable to SUN360). It accepts 32 RGB images of size 224×224 cropped from 8 frames and outputs the lighting estimate through the alignment and averaging process. We borrow the core structure of the transformer from [43] and carefully determine the number of layers,

number of heads, hidden size, and MLP size as 4, 4, 512, and 1024, respectively, under extensive experiments. The dropout rate was 0.2.

We train our model and test its performance on the SUN360 and KITTI datasets separately (see Table 4.2). In detail, we empirically trained our lighting estimation network for 118 and 131 epochs for the SUN360 and KITTI datasets using early stopping. The training was initiated with the AdamW optimizer [80] using a learning rate of 1×10^{-5} and the batch size was 8. It took 61.1 and 34.3 hours on a single Nvidia RTX 3090 GPU. Prediction on a single sequence of 32 images takes 90 ms. Our spatio-temporal aggregation model is examined on 1000 unobserved SUN360 sequences and 432 KITTI sequences.

4.4 Results

4.4.1 Four-Stage Approach

Sun Direction

We evaluate the angular errors of the spatially aggregated sun direction estimates on the SUN360 test set. At first, single image lighting estimation results are gathered using [53], [55], [57], and our method. Then we compensate the camera angles and apply our spatial aggregation step on the subimages to acquire the spatially combined estimate for each panorama. The explicit spatial aggregation step involves two additional hyperparameters: the contamination ratio and the mean-shift kernel width. We found those parameters to be insensitive to different data sets and kept the same values in all our experiments. The *contamination ratio* is set to 0.5 because we assume the estimations with angular errors larger than an octant ($22.5°$) as outliers, which is roughly $50\,\%$ of the data for our method when observing Fig. 4.11. As a result, we apply the mean shift algorithm on $50\,\%$ potential inliers among the total observations.

Fig. 4.8 illustrates the cumulative angular errors of the four methods. Since the previous methods were trained with only the SUN360 training set, due to the characteristics of their networks (requiring ground truth exposure and turbidity information which are lacked in the KITTI dataset), we also report our method's performance when it was trained only on SUN360 (see *Ours, SUN360* in Fig. 4.8). Our method performs better than the previous techniques even with the same training set. The detailed quantitative comparison is presented in Fig. 4.11. Note that all methods are trained and tested with subimages instead of full images.

For the KITTI dataset, we can further extend the lighting estimation to the temporal domain. Although the dataset provides the ground truth ego-motion, we calculated it using [34] to generalize our approach. The mean angular error of the estimated camera rotation using

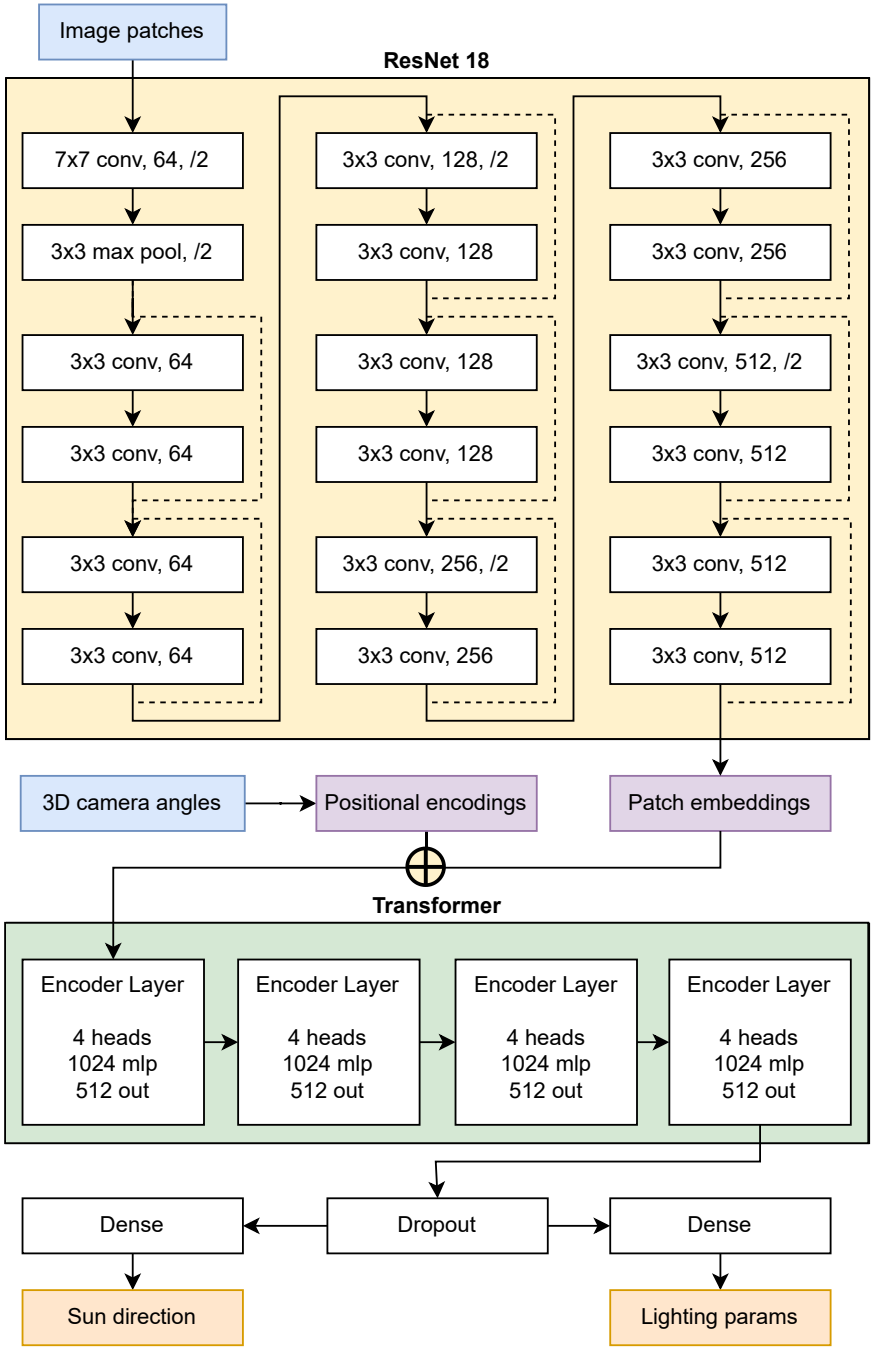


Figure 4.7. The proposed lighting estimation model. The features of the input image patches are extracted through the ResNet18 [6] network. We generate orientation-invariant positional encodings from the given 3D camera angles and add them (denoted as $\oplus$) to the patch embeddings. Our transformer network then aggregates the observations and outputs the estimated sun direction and lighting parameters of the sequence. Note that the right-side dense layer is omitted for the KITTI dataset.

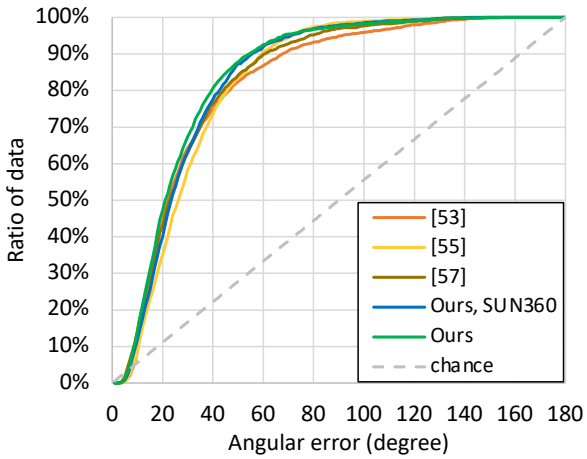


Figure 4.8. The cumulative angular error for spatially aggregated sun direction estimates on the SUN360 test set. *Ours, SUN360* indicates our results when the network was only trained with the SUN360 dataset.

Sequence	Single	Spatial	Spatiotemporal
(a)	13.43	6.76	3.54
(b)	26.06	7.81	6.87
(c)	34.68	24.83	13.17
(d)	23.03	10.04	3.27

Table 4.3. Angular errors of each aggregation step (from left to right: single image (baseline), spatial aggregation, spatio-temporal aggregation). Sequences correspond to Fig. 4.9.

the default parameters was $1.01°$ over the five test sequences. We plotted the sun direction estimates of each step in our pipeline for four (out of five) test sequences in Fig. 4.9. Note that in the plots all predictions are registered to a common coordinate frame using the estimated camera ego-motion. Individual estimates of the subimages are shown with gray dots. Our spatial aggregation process refines the noisy observations using outlier removal and mean shift (black dots). Those estimates for each frame in a sequence are finally combined in the temporal aggregation step (denoted with the green dot). The ground truth direction is indicated by the red dot. Using the spatio-temporal filtering, the mean angular error over the five test sequences recorded $7.68°$, which is a reduction of $69.94\,\%$ ($25.56°$ for single image based estimation). A quantitative evaluation of the performance gain for each aggregation step is presented in Table 4.3.

Our model's stability is better understood with a virtual object augmentation application as shown in Fig. 4.10. Note that other lighting parameters, such as the sun's intensity are manually determined. When the lighting conditions are estimated from only a single image on each frame, the virtual objects' shadows are fluctuating compared to the ground truth results. The artifact is less visible on our spatial aggregation results and entirely removed

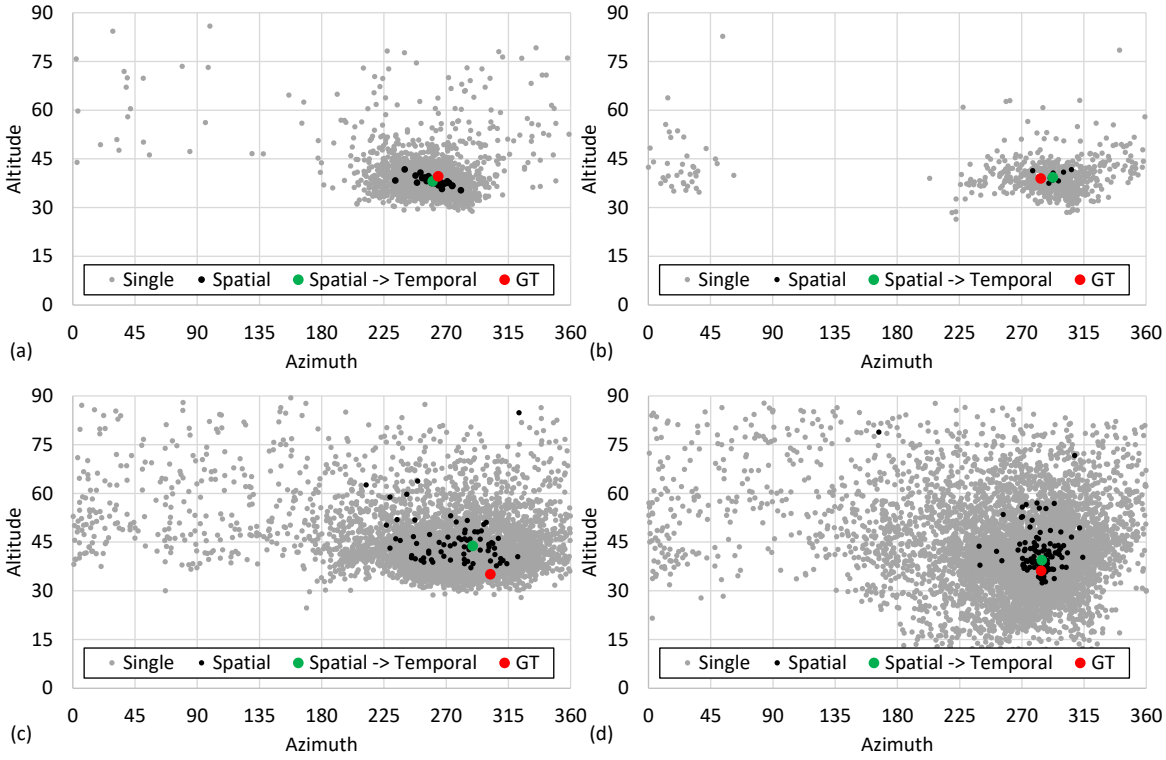


Figure 4.9. Scatter plots representing sun direction estimates of individual subimages and the results of two aggregation steps. Each graph corresponds to an image sequence in the KITTI test set. Despite numerous outliers in the raw observations (the gray dots), our two-step aggregation determines the video's lighting condition with small margins to the ground truth sun direction (the black dots for spatial aggregation and the green dot for spatio-temporal aggregation). Angular errors for our spatio-temporal filtering results are (a) 3.54 (b) 6.87 (c) 13.17 and (d) 3.27 degrees.

after applying the spatio-temporally aggregated lighting condition.

Ablation Study

The performance gain of the spatial aggregation process is thoroughly analyzed by breaking down the individual filtering steps on the SUN360 test set. Fig. 4.11 shows the cumulative angular error for the raw observations and compares the four lighting estimation methods with four different aggregation strategies:

- *Single*: unprocessed individual observations,
- *Mean all*: mean of all estimates from each panorama,
- *Mean inliers*: mean of inlier estimates,

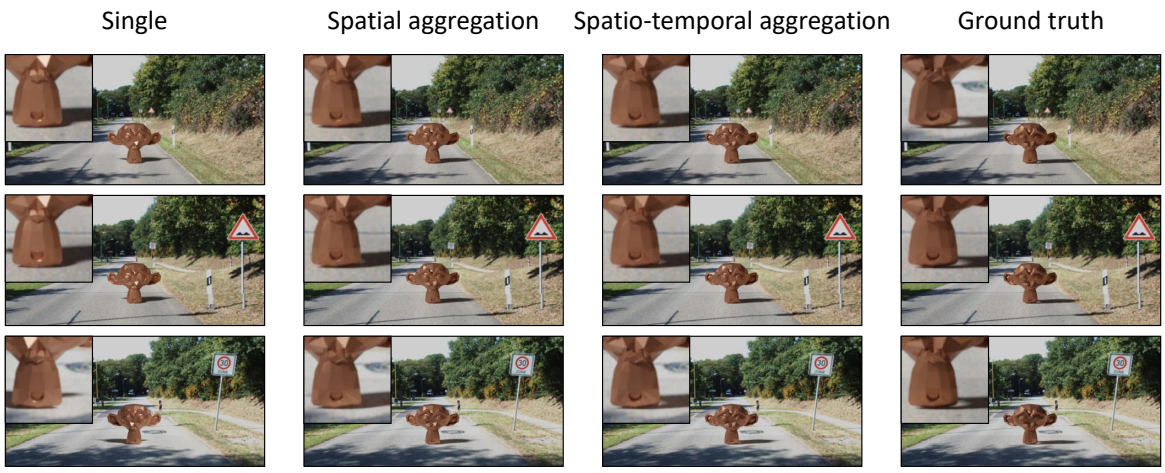
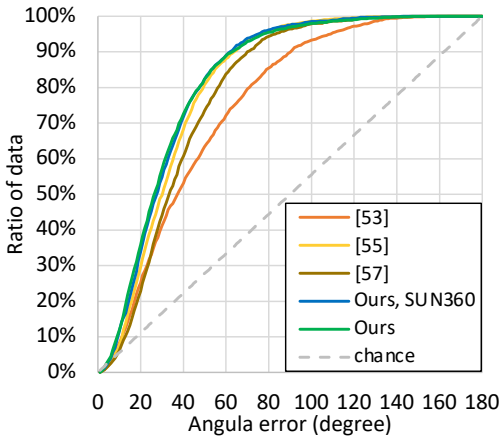
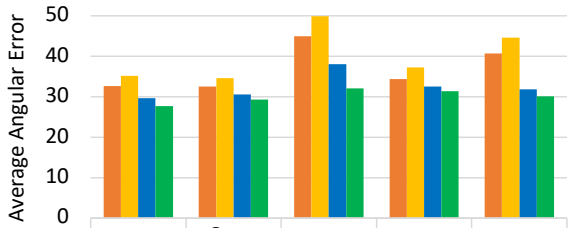


Figure 4.10. Demonstration of a virtual augmentation application. Fluctuations in the shadow of the augmented object decrease as the estimates are refined through our pipeline. After applying the spatio-temporal filtering, the results are fully stabilized and almost indistinguishable from the ground truth. Please also refer to the augmented video in the supplementary material.

- *Meanshift*: mean shift result of inlier estimates.





	Ours	Ours, SUN360	[53]	[55]	[57]
Single	32.64	**32.57**	44.91	34.32	40.75
Mean All	35.11	**34.61**	49.94	37.21	44.61
Mean Inlier	**29.61**	30.58	38.03	32.53	31.80
Meanshift	**27.63**	29.26	32.04	31.32	30.13

Figure 4.11. (*left*) The cumulative angular error for the *single* estimates on the SUN360 test set. (*right*) Comparing average angular error for three methods with different spatial aggregation strategies. Our method achieved the best result when the mean shift is applied to the inliers. We outperform previous methods even without the KITTI dataset.

As illustrated in Fig. 4.11, the average angular error of each method is decreased by at most 10 degrees after applying the proposed spatial aggregation. This result demonstrates our method's generality, showing that it can increase the accuracy of any lighting estimation

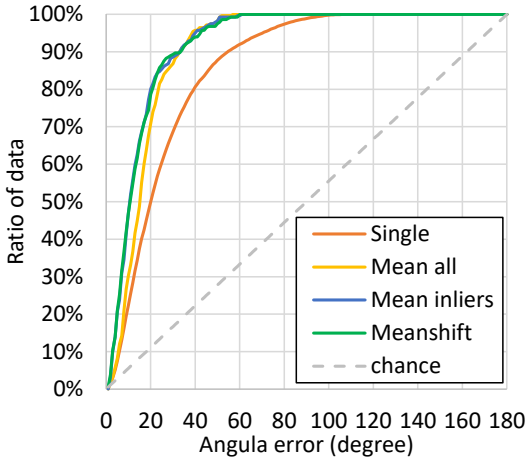


Figure 4.12. The cumulative angular error on the KITTI test set with different spatial aggregation strategies. The best result is recorded when the mean shift result of the inlier estimates is utilized.

method. We observe a slight increase in the average error for the *Mean all* metric due to the outlier observations. A similar analysis is done for the KITTI dataset with only our method. The cumulative angular error graphs for the four steps are presented in Fig. 4.12.

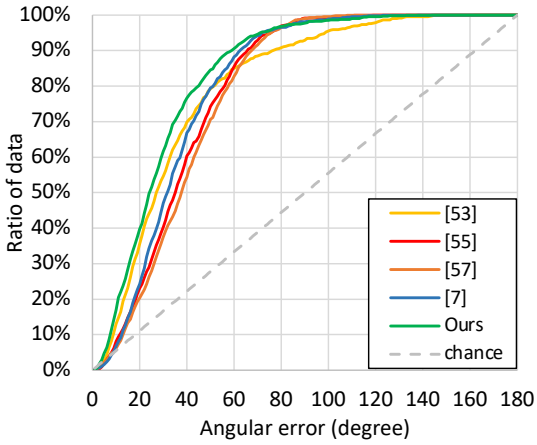
4.4.2 End-to-End Approach

Sun Direction

We evaluate the angular errors of the spatio-temporally aggregated sun direction estimates on the SUN360 test sequences. Since other single image-based lighting estimation methods [53, 55, 57] are not capable of conducting spatio-temporal aggregation, the median of the estimates over each sequence is utilized. On top of that, we compare our method with the spatio-temporal aggregation pipeline proposed in [7]. The hyperparameters required for our previous method are determined in the same way as described in [7].

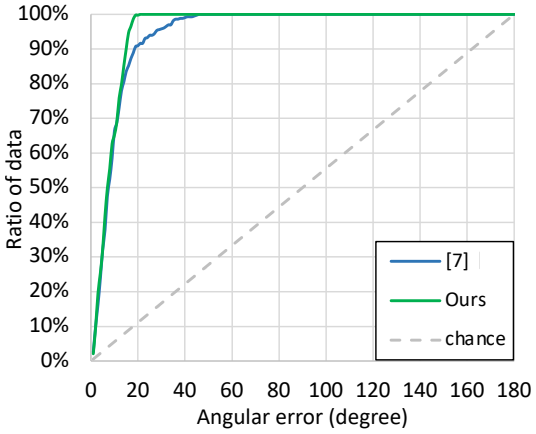
Fig. 4.13 illustrates the cumulative angular errors of the five methods trained and tested on the SUN360 dataset. We present the outcomes of three single image based approaches along with the results of two spatio-temporal aggregation methods. Our spatio-temporal attention method shows a noticeable margin compared to the state-of-the-art.

We also performed a similar comparison on the KITTI dataset (see Fig. 4.14). On this dataset, however, we compare our method only with [7] due to the lack of ground truth information such as exposure and turbidity which are required for other previous works. Although the dataset provides the ground truth ego-motion required for the alignment step, we calculated it using [34] to generalize our approach. The mean angular error of the estimated camera rotation using the default parameters was 1.01° over the five test scenes.



	Median	Mean	Min	Max	SD
[53]	27.00	35.39	1.27	161.03	0.3325
[55]	35.01	37.36	0.84	**118.10**	0.1895
[57]	37.75	39.12	**0.21**	126.65	0.1915
[7]	31.66	35.20	0.33	137.57	0.1297
Ours	**24.12**	**29.41**	0.78	143.47	**0.0569**

Figure 4.13. The cumulative angular error and the statistics of the sun direction estimates on the SUN360 test set. *[7]* and *Ours* are showing the spatiotemporal aggregation results. For a fair comparison, angular errors of other methods are measured upon the median of the estimates made on single images. The proposed method outperforms other methods with a noticeable margin.



	Median	Mean	Min	Max	SD
[7]	7.42	9.62	**0.23**	45.93	0.0838
Ours	**7.04**	**7.96**	0.55	**20.42**	**0.0142**

Figure 4.14. The cumulative angular error and the statistics on the KITTI test set. Our method performs slightly better than [7] while recording a noticeable small maximum angular error of $20.42°$.

Using the proposed spatio-temporal attention method, the mean angular error over the 432 test sequences recorded $7.96°$, which is marginally better than $9.62°$ of [7].

We plotted the individual sun direction estimates and their aggregation results using our methods and [7] in Fig. 4.15. Note that in the plots all predictions are registered to a common coordinate frame using the estimated camera ego-motion. Individual estimates of the subimages are shown with lighter color dots. The single image estimation of [7] was performed individually and resulted in independent noisy estimates which were aggregated by statistical post-processing. Unlike them, our estimates are jointly predicted and therefore tend to cluster tightly around their mean rendering any statistical post-processing redundant.

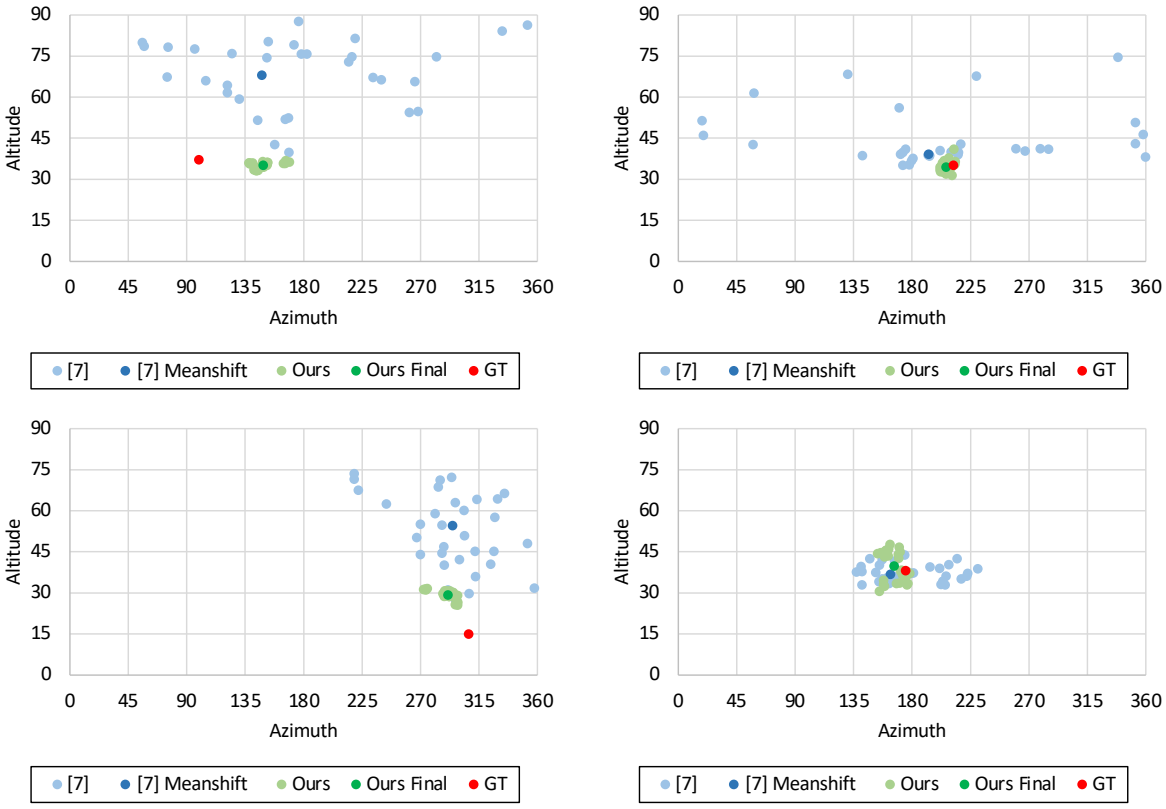


Figure 4.15. Scatter plots representing sun direction estimates of individual subimages and the spatiotemporal aggregation result. Each plot corresponds to an image sequence of 8 frames in (*left*) the SUN360 and (*right*) the KITTI test sets. The spatio-temporal aggregation proposed in [7] finds the highest point density among the inliers treating the estimates as independent sample. On the contrary, individual estimates of our method form a tight group due to the spatio-temporal attention.

The mean standard deviation of sun direction estimations also demonstrates our model's capability for coherent estimation (see Fig. 4.13 and Fig. 4.14). Compared to other methods, we recorded 2 to 6 times lower mean standard deviation. This behavior comes from the spatio-temporal attention from our transformer network. We contend that the network tries to output a set of predictions that can explain the lighting condition of the given sequence, rather than predicting each subimage's lighting condition individually. Furthermore, this characteristic supports our decision to average all estimates to obtain the final estimate of the sequence.

Other Lighting Parameters

As described earlier, the remaining *Lalonde-Matthews* model's parameters are only estimated for the SUN360 dataset. We present the root mean squared errors of [53], [55], and

	$\mathbf{w}_{sun}$	$\mathbf{w}_{sky}$	κ	β	t
[53]	-	-	-	-	1.0869
[55]	0.3680	0.1083	0.1817	9.7960	1.1994
Ours	**0.2810**	**0.0833**	**0.1201**	**6.9778**	**0.9510**

Table 4.4. RMSE of the estimated parameters on the SUN360 test set

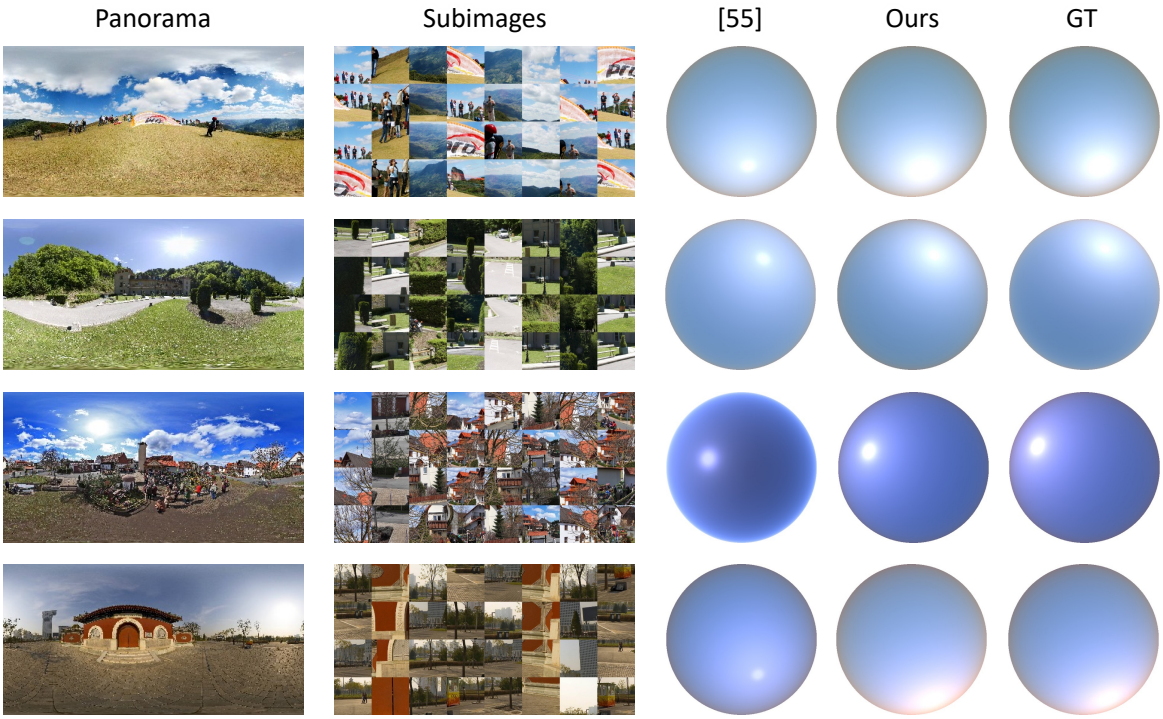


Figure 4.16. Qualitative comparison on the estimated parameters of the *Lalonde-Matthews* model. Our methods aggregates information obtained from the subimages of a synthetic sequence and provides plausible outcomes on various lighting conditions.

ours in Table. 4.4. Note that [53] only delivers the RMSE for turbidity, because it is based on a different lighting model. Our method demonstrated outstanding performance for all five items. We also provide a qualitative evaluation on the full *Lalonde-Matthews* model in Fig. 4.16. Each hemispherical texture is generated using the estimated/ground truth parameters.

The stability of our model is better understood with a virtual object augmentation application, as shown in Fig. 4.17. Note that other lighting parameters, such as the sun's intensity, are manually determined and equally applied for the single image estimation method and [7]. When the lighting conditions are estimated from only a single image on each frame, the vir-

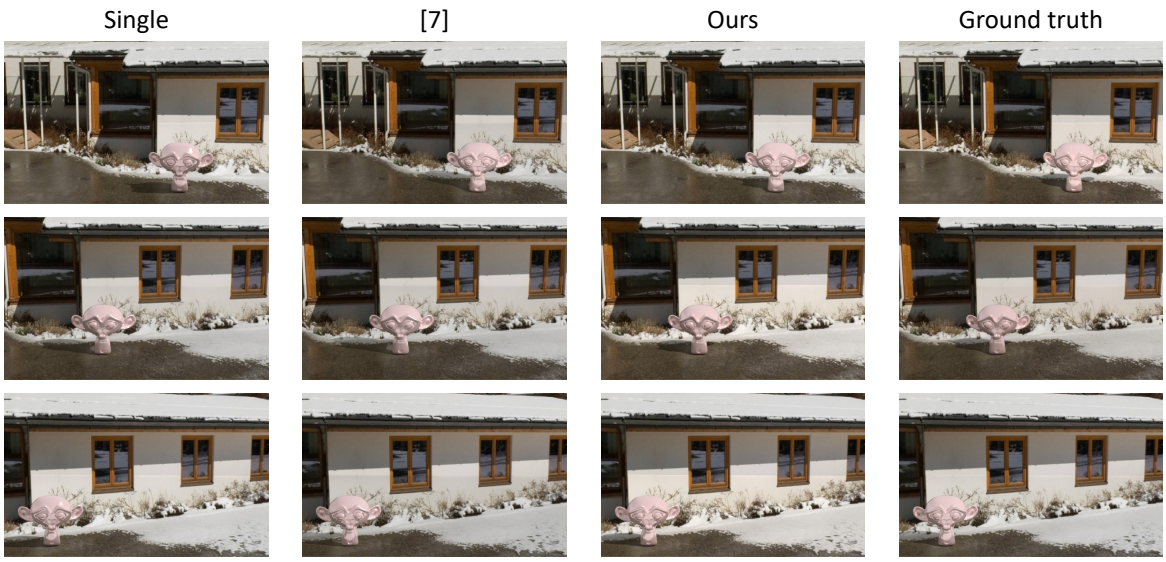


Figure 4.17. Virtual Augmentations: Fluctuations in the shadow of the augmented object are strongly visible when the sun direction is estimated individually. Our spatio-temporal method [7] achieves more stable results. The proposed learned aggregation results in even better quality, almost indistinguishable from the ground truth.

tual objects' shadows are fluctuating compared to the ground truth results. The artifact is almost entirely removed and the augmented object's appearance is almost identical to the ground truth after applying the spatio-temporally aggregated lighting condition based on the *Lalonde-Matthews* model.

Ablation Study

We perform a series of ablations for our chosen losses, positional encoding and the number of patches for our model. Ablations are done on the SUN360 test set and we compare angular error statistics.

Loss Function Table 4.5 shows the angular error statistics for different loss term combinations. The L_{cosine} metric was set as the default loss function as it dominantly drives the training. Best performance can be achieved by using all loss terms together.

Positional Encoding We investigate the benefit of our newly proposed orientation-invariant positional encoding by comparing it to the standard sinusodial encoding introduced in [42]. The results in Table. 4.6 show, that our task-specific encoding gives greater performance over the standard one or using none at all.

L_{cosine}	L_{norm}	L_{hemi}	Median	Mean	Min	Max
✓			26.20	31.47	1.24	157.98
✓	✓		25.00	30.59	**0.29**	157.96
✓		✓	25.04	30.94	0.51	157.65
✓	✓	✓	**24.12**	**29.41**	0.78	**143.47**

Table 4.5. Ablation study with loss functions on the SUN360 test set

	Median	Mean	Min	Max
None	35.56	37.99	1.30	157.11
Standard	27.42	32.06	**0.55**	165.64
Ours	**24.12**	**29.41**	0.78	**143.47**

Table 4.6. Ablation study with positional encoding schemes on the SUN360 test set

Frames	Subimages	Median	Mean	Min	Max
4	4	25.83	31.31	**0.66**	155.42
8		**24.12**	**29.41**	0.78	**143.47**
12		24.62	30.33	0.97	151.44
16		25.91	31.02	1.11	160.40
8	2	24.87	30.82	**0.58**	160.38
	4	**24.12**	**29.41**	0.78	**143.47**
	6	24.53	30.60	0.80	173.33
	8	25.55	31.29	0.91	152.70

Table 4.7. Ablation study with hyperparameters on the SUN360 test set

Patch Sequence In these experiments, we ablate the number and choice of patches given to the aggregation transformer. By changing the number of frames and number of spatial patches per image, we compare different temporal-spatial patch variations. The results in Table 4.7 show that there is a sweet spot for the length of the temporal sequence and the number of patches per image. We achieve the best performance by choosing a sequence of 8 images and 4 patches per image, resulting in a sequence length of 32. Increasing the sequence length seems to hurt the model performance at a certain point. We believe that this could be due to the limited model capacity and plan to experiment with larger networks in the future.

4.5 Conclusion

This paper introduced two innovative approaches to outdoor lighting estimation, each contributing uniquely to advancements in the field. The first approach developed a single image lighting estimation method, significantly enhanced through spatial and temporal aggregation. This approach demonstrated state-of-the-art performance in outdoor lighting estimation for individual image sequences and proved its adaptability by effectively utilizing $360°$ panoramas and wide-view images. Its capability for spatio-temporal aggregation not only showed versatility in processing various image types but also opened pathways for applying this methodology to a wide range of globally shared scene information gathering methods.

Building on the foundations laid by the first approach, the second approach presented a holistic sequence-wise lighting estimation method using spatio-temporal attention with transformers. This advanced model, achieving state-of-the-art performance, addressed and improved upon some of the limitations found in the first approach. It introduced a comprehensive end-to-end pipeline, eliminating the need for additional post-processing steps. Furthermore, the implementation of the advanced Lalonde-Matthews sun-sky lighting model in the second approach enhanced the detail and accuracy of the lighting estimates, surpassing the capabilities of the first approach.

The integration of the Lalonde-Matthews model in the second approach marks a significant step forward from the first, catering to a broader range of applications, particularly in areas demanding high-precision lighting information, such as photorealistic virtual object augmentation in image sequences. This progression from the first to the second approach underlines the continual evolution and improvement in outdoor lighting estimation techniques.

Despite these advancements, future research directions remain open and promising. Scaling both the model and data in the second approach to explore the limits of attention-based spatio-temporal aggregation for lighting estimation presents an exciting avenue for

further exploration. Additionally, integrating this advanced methodology into reconstruction pipelines such as SLAM, where accurate lighting direction and shadow-casting can aid in camera estimation, is another promising direction. Exploring varied sampling methods, such as selecting consecutive frames and experimenting with the number of frames and their arrangement, offers potential for further refining and enhancing the model's performance.

In conclusion, both approaches make significant contributions to the field of lighting estimation, each building upon the other's strengths and paving the way for future innovations. These developments not only enhance our understanding and capabilities in image processing but also open new possibilities in augmented reality and beyond.

Chapter 5

Conclusions and Outlook

5.1 Conclusions

This dissertation, titled "Lighting Estimation in Outdoor Scenes" has made strides in advancing the field of outdoor lighting estimation. The journey, articulated through two key chapters, began with an exploration into intrinsic image decomposition and presented the development of an end-to-end spatio-temporal lighting estimation pipeline.

Chapter 3 delved into intrinsic image decomposition using a U-Net architecture coupled with a Multi-Layer Perceptron (MLP) network. This phase aimed to extract lighting direction information from images, positing that the latent vector at the U-Net's deepest level encoded critical lighting conditions. The overall structure is designed using a Siamese network configuration with triplet loss and trained predominantly on synthetic scenes rendered in Blender. This approach offered a deeper understanding of the decomposition process, although it faced its own set of challenges, particularly in handling complex, realistic scenes.

Chapter 4 marked an advancement in lighting estimation techniques. It presented a comprehensive transition from the initial single-image-based approach with subsequent post-processing to an innovative end-to-end model. This model leveraged a Transformer architecture to streamline the estimation process, eliminating the need for intricate hyperparameter tuning and enhancing the efficiency of the estimation process. The model's incorporation of a novel handcrafted positional encoding mechanism tailored for spatio-temporal aggregation and the adoption of the sophisticated Lalonde-Matthews outdoor illumination model underscored its novelty and effectiveness. The chapter highlighted the successful application of an attention-based model for lighting estimation and its superiority over existing state-of-the-art methods.

In summary, this dissertation has not only contributed to the realm of precise outdoor lighting estimation but also demonstrated the potential of deep learning in addressing intricate, ill-posed problems in computer vision. These contributions pave the way for more

accurate and realistic interpretations of outdoor scenes and have broad implications for various applications in computer vision and beyond.

5.2 Outlook

The trajectory of computer vision research, particularly in relation to lighting estimation, opens several promising paths for future exploration. This dissertation lays the groundwork for these explorations, highlighting areas ripe for development and innovation.

Performance Improvement and Domain Adaptation Enhancing the performance of the lighting estimation method, especially in terms of domain adaptation, is a vital future endeavor. Exploring domain transfer techniques or linking with Large Language Models (LLMs) could mitigate the challenges posed by domain gaps. Further research could focus on developing adaptive algorithms capable of adjusting to various environmental conditions, thereby broadening the applicability of lighting estimation models across diverse datasets.

Estimating Spatially Varying Lighting Conditions Investigating the estimation of spatially varying lighting conditions opens new horizons for more nuanced and realistic lighting models. This line of research could delve into how different environmental elements interact with light, leading to more accurate simulations in virtual environments. Such advancements would significantly benefit applications in virtual reality (VR), gaming, and cinematography, where spatially accurate lighting plays a critical role in creating immersive experiences.

Intrinsic Image Decomposition The pursuit of intrinsic image decomposition remains a fertile area for exploration, despite the challenges encountered. Employing sim2real techniques, such as rendering photorealistic images from 3D scenes using advanced diffusion models, could provide the necessary data for training robust decomposition networks. This approach could also offer new insights into the complex interplay between lighting, material properties, and perception, enhancing the understanding and application of intrinsic decomposition in various fields.

Applications in Computer Vision and Generative Models Extending the lighting estimation method to other facets of computer vision holds substantial potential. By accurately predicting lighting conditions, algorithms for shadow detection, material estimation, and SLAM could see significant improvements in accuracy and realism. In the realm of generative models, integrating lighting estimation could revolutionize image synthesis, enabling

the creation of photorealistic scenes with dynamically adjusted lighting that reflects real-world conditions.

Comprehensive Scene Understanding and AR Applications The advancement in lighting estimation aligns with the overarching objective of achieving comprehensive scene understanding. This research direction is particularly relevant for augmented reality (AR), where the ability to blend virtual objects seamlessly into real-world environments hinges on accurate lighting estimation. Future developments in this area could lead to more realistic and immersive AR experiences, propelling the field towards new heights of interactivity and engagement.

Estimating Lighting Conditions with Multiple Light Sources Venturing into the realm of multi-source lighting estimation presents a significant advancement for the field. This research direction would cater to scenarios where lighting is not solely dependent on natural light, such as indoor environments or night-time settings. Developing methods to accurately predict the interaction and cumulative effect of multiple light sources would substantially enhance the realism and applicability of lighting models. It could lead to breakthroughs in areas like night-time photography enhancement, indoor scene rendering, and even in safety-critical applications like nighttime navigation for autonomous vehicles. Exploring this area could also shed light on the complex dynamics of light propagation and reflection in varied environments, contributing to a more holistic understanding of scene illumination.

Bibliography

[1] Theia. *Theia Vision Library*.

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

[3] Jianxiong Xiao, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Recognizing scene viewpoint using panoramic place representation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2695–2702. IEEE, 2012.

[4] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. *arXiv preprint arXiv:1611.05431*, 2016.

[5] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[7] Haebom Lee, Robert Herzog, Jan Rexilius, and Carsten Rother. Spatiotemporal outdoor lighting aggregation on image sequences. In *DAGM German Conference on Pattern Recognition*, pages 343–357. Springer, 2021.

[8] James T Kajiya. The rendering equation. In *Proceedings of the 13th annual conference on Computer graphics and interactive techniques*, pages 143–150, 1986.

[9] Matt Pharr, Wenzel Jakob, and Greg Humphreys. *Physically based rendering: From theory to implementation*. MIT Press, 2023.

[10] Turner Whitted. An improved illumination model for shaded display. *Communications of the ACM*, 23(6):343–349, 1980.

[11] Cindy M Goral, Kenneth E Torrance, Donald P Greenberg, and Bennett Battaile. Modeling the interaction of light between diffuse surfaces. *SIGGRAPH '84 Proceedings*, 18(3):213–222, 1984.

[12] Ruo Zhang. Shape from shading: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(8):690–706, 1999.

[13] Paul E Debevec. Rendering synthetic objects into real scenes: Bridging traditional and image-based graphics with global illumination and high dynamic range photography. *SIGGRAPH '98 Proceedings*, 1998.

[14] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003.

[15] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, 2000.

[16] Richard Szeliski. *Computer Vision: Algorithms and Applications*. Springer Science & Business Media, 2010.

[17] Erik Reinhard, Greg Ward, Sumanta Pattanaik, and Paul Debevec. *High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting*. Morgan Kaufmann, 2010.

[18] James D Foley, Andries van Dam, Steven K Feiner, and John F Hughes. *Computer Graphics: Principles and Practice*. Addison-Wesley Professional, 1996.

[19] Jean-François Lalonde and Iain Matthews. Lighting estimation in outdoor image collections. In *2014 2nd International Conference on 3D Vision*, volume 1, pages 131–138. IEEE, 2014.

[20] Fred E Nicodemus, John C Richmond, J J Hsia, Ira W Ginsberg, and Thomas Limperis. *Directional reflectance and emissivity of an opaque surface*. US Department of Commerce, National Bureau of Standards, 1977.

[21] Jos Stam. Diffraction shaders. *SIGGRAPH '99 Proceedings*, pages 101–110, 1999.

[22] James F Blinn. Simulation of wrinkled surfaces. *SIGGRAPH '78 Proceedings*, 12(3):286–292, 1978.

[23] Henrik Wann Jensen, Stephen R Marschner, Marc Levoy, and Pat Hanrahan. A practical model for subsurface light transport. *SIGGRAPH '01 Proceedings*, pages 511–518, 2001.

[24] Robert L Cook. Shade trees. In *Proceedings of the 11th annual conference on Computer graphics and interactive techniques*, pages 223–231, 1984.

[25] Jules Bloomenthal, Brian Wyvill, Chandrajit L Bajaj, James F Blinn, Marie-Paule Cani, Alyn Rockwood, Geoff Wyvill, and Stanley Osher. *Introduction to Implicit Surfaces*. Morgan Kaufmann, 1997.

[26] Bui Tuong Phong. Illumination for computer generated pictures. *Communications of the ACM*, 18(6):311–317, 1975.

[27] Philip Dutre, Philippe Bekaert, and Kavita Bala. *Advanced global illumination*. CRC Press, 2018.

[28] Henrik Wann Jensen. Global illumination using photon maps. In *Rendering Techniques*, pages 21–30. Springer, 1996.

[29] Peter-Pike Sloan, Jan Kautz, and John Snyder. Precomputed radiance transfer for real-time rendering in dynamic, low-frequency lighting environments. *ACM Transactions on Graphics (TOG)*, 21(3):527–536, 2002.

[30] Sébastien Lagarde and Charles de Rousiers. Moving frostbite to pbr, 2014. Blog post.

[31] David J Crandall, Andrew Owens, Noah Snavely, and Daniel P Huttenlocher. Discrete-continuous optimization for large-scale structure from motion. In *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3001–3008. IEEE, 2011.

[32] Marc Pollefeys, Luc Van Gool, Maarten Vergauwen, Frank Verbiest, Kurt Cornelis, Jan Tops, and Reinhard Koch. Visual modeling with a hand-held camera. *International Journal of Computer Vision*, 59(3):207–232, 2004.

[33] Bill Triggs, Philip F McLauchlan, Richard I Hartley, and Andrew W Fitzgibbon. Bundle adjustment—a modern synthesis. In *International workshop on vision algorithms*, pages 298–372. Springer, 1999.

[34] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4104–4113, 2016.

[35] Johannes Lutz Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, pages 501–518. Springer, 2016.

[36] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[37] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[38] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[39] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

[40] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[41] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[43] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[44] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[45] Blender Foundation. *Blender - a 3D modelling and rendering package*.

[46] Harry Barrow, J Tenenbaum, A Hanson, and E Riseman. Recovering intrinsic scene characteristics. *Comput. vis. syst*, 2(3-26):2, 1978.

[47] Takuya Narihira, Michael Maire, and Stella X Yu. Direct intrinsics: Learning albedo-shading decomposition by convolutional regression. In *Proceedings of the IEEE international conference on computer vision*, pages 2992–2992, 2015.

[48] Tom Haber, Christian Fuchs, Philippe Bekaer, Hans-Peter Seidel, Michael Goesele, and Hendrik PA Lensch. Relighting objects from image collections. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 627–634. IEEE, 2009.

[49] Jean-François Lalonde, Alexei A Efros, and Srinivasa G Narasimhan. Estimating the natural illumination conditions from a single outdoor image. *International Journal of Computer Vision*, 98(2):123–145, 2012.

[50] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[51] Yongjie Zhu, Yinda Zhang, Si Li, and Boxin Shi. Spatially-varying outdoor lighting estimation from intrinsics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12834–12842, 2021.

[52] Haebom Lee, Christian Homeyer, Robert Herzog, Jan Rexilius, and Carsten Rother. Spatio-temporal outdoor lighting aggregation on image sequences using transformer networks. *International Journal of Computer Vision*, 131(4):1060–1072, 2023.

[53] Yannick Hold-Geoffroy, Kalyan Sunkavalli, Sunil Hadap, Emiliano Gambaretto, and Jean-François Lalonde. Deep outdoor illumination estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7312–7321, 2017.

[54] Yannick Hold-Geoffroy, Akshaya Athawale, and Jean-François Lalonde. Deep sky modeling for single image outdoor lighting estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6927–6935, 2019.

[55] Xin Jin, Pengyue Deng, Xinxin Li, Kejun Zhang, Xiaodong Li, Quan Zhou, Shujiang Xie, and Xi Fang. Sun-sky model estimation from outdoor images. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–12, 2020.

[56] Jinsong Zhang, Kalyan Sunkavalli, Yannick Hold-Geoffroy, Sunil Hadap, Jonathan Eisenman, and Jean-François Lalonde. All-weather deep outdoor lighting estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10158–10166, 2019.

[57] Kejun Zhang, Xinxin Li, Xin Jin, Biao Liu, Xiaodong Li, and Hongbo Sun. Outdoor illumination estimation via all convolutional neural networks. *Computers & Electrical Engineering*, 90:106987, 2021.

[58] Housheng Wei, Yanli Liu, Guanyu Xing, Yanci Zhang, and Wenjia Huang. Simulating shadow interactions for outdoor augmented reality with rgbd data. *IEEE Access*, 7:75292–75304, 2019.

[59] Reza Mahjourian, Martin Wicke, and Anelia Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5667–5675, 2018.

[60] Tom Van Dijk and Guido C. H. E. de Croon. How do neural networks see depth in single images? In *Proceedings of the IEEE international conference on computer vision*, pages 2183–2191, 2019.

[61] Vignesh Prasad and Brojeshwar Bhowmick. Sfmlearner++: Learning monocular depth & ego-motion using meaningful geometric constraints. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2087–2096. IEEE, 2019.

[62] Hasan Balcı and Uğur Güdükbay. Sun position estimation and tracking for virtual object placement in time-lapse videos. *Signal, Image and Video Processing*, 11(5):817–824, 2017.

[63] Peter Kán and Hannes Kaufmann. Deeplight: light source estimation for augmented reality using deep learning. *The Visual Computer*, 35(6-8):873–883, 2019.

[64] Claus B Madsen and Brajesh B Lal. Outdoor illumination estimation in image sequences for augmented reality. *GRAPP*, 11:129–39, 2011.

[65] Kevin Karsch, Varsha Hedau, David Forsyth, and Derek Hoiem. Rendering synthetic objects into legacy photographs. *ACM Transactions on Graphics (TOG)*, 30(6):1–12, 2011.

[66] Boun Vinh Lu, Tetsuya Kakuta, Rei Kawakami, Takeshi Oishi, and Katsushi Ikeuchi. Foreground and shadow occlusion handling for outdoor augmented reality. In *2010 IEEE International Symposium on Mixed and Augmented Reality*, pages 109–118. IEEE, 2010.

[67] Xin Jin, Xing Sun, Xiaokun Zhang, Hongbo Sun, Ri Xu, Xinghui Zhou, Xiaodong Li, and Ruijun Liu. Sun orientation estimation from a single image using short-cuts in dcnn. *Optics & Laser Technology*, 110:191–195, 2019.

[68] Wei-Chiu Ma, Shenlong Wang, Marcus A Brubaker, Sanja Fidler, and Raquel Urtasun. Find your way by observing the sun and other semantic cues. In *2017 IEEE In-*

ternational Conference on Robotics and Automation (ICRA), pages 6292–6299. IEEE, 2017.

[69] Yanli Liu and Xavier Granier. Online tracking of outdoor lighting variations for augmented reality with moving cameras. *IEEE Transactions on visualization and computer graphics*, 18(4):573–580, 2012.

[70] Claus B Madsen, Moritz Störring, Tommy Jensen, Mikkel S Andersen, and Morten F Christensen. Real-time illumination estimation from image sequences. In *Proceedings: 14th Danish Conference on Pattern Recognition and Image Analysis, Copenhagen, Denmark*, pages 1–9, 2005.

[71] Ibrahim Reda and Afshin Andreas. Solar position algorithm for solar radiation applications. *Solar energy*, 76(5):577–589, 2004.

[72] Yuan Xiong, Hongrui Chen, Jingru Wang, Zhe Zhu, and Zhong Zhou. Dsnet: Deep shadow network for illumination estimation. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pages 179–187. IEEE, 2021.

[73] Lukas Hosek and Alexander Wilkie. An analytic model for full spectral sky-dome radiance. *ACM Transactions on Graphics (TOG)*, 31(4):1–9, 2012.

[74] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422. IEEE, 2008.

[75] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[76] Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 24(5):603–619, 2002.

[77] Arcot J Preetham, Peter Shirley, and Brian Smits. A practical analytic model for daylight. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 91–100, 1999.

[78] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.

[79] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[80] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[81] Richard Perez, Robert Seals, and Joseph Michalsky. All-weather model for sky luminance distribution—preliminary configuration and validation. *Solar energy*, 50(3):235–245, 1993.

Appendix A

Preetham Sky Model

Most of the text in this appendix is directly taken from the text and appendix of [77] to explain how f_P in 4.2.2 is defined.

The original Preetham sky model is defined in the CIE xyY color space:

$$\begin{aligned} x &= x_z \mathcal{F}_x(\theta, \gamma)/\mathcal{F}_x(0, \theta_s) \\ y &= y_z \mathcal{F}_y(\theta, \gamma)/\mathcal{F}_y(0, \theta_s) \\ Y &= Y_z \mathcal{F}_Y(\theta, \gamma)/\mathcal{F}_Y(0, \theta_s), \end{aligned} \tag{A.1}$$

where x_z, y_z, and Y_z are the xyY color values of the zenith and θ_s is the zenith angle of the sun. The function $\mathcal{F}$ is the Perez et al.'s sky luminance distribution model [81]:

$$\mathcal{F}(\theta, \gamma) = (1 + Ae^{B/\cos\theta})(1 + Ce^{D\gamma} + E\cos^2\gamma), \tag{A.2}$$

where A, B, C, D, and E are the distribution coefficients, θ is the angle between the viewing direction and the zenith, and γ is the angle between the viewing direction and the sun.

Absolute value of zenith luminance in K cd m^{-2}:

$$Y_z = (4.0453T - 4.9710)\tan\chi - 0.2155T + 2.4192, \tag{A.3}$$

where $\chi = (\frac{4}{9} - \frac{T}{120})(\pi - 2\theta_s)$ and T is the turbidity.

Zenith chromaticity (x_z, y_z):

$$\begin{aligned} x_z &= \begin{bmatrix} T^2 & T & 1 \end{bmatrix} \begin{bmatrix} 0.0017 & -0.0037 & 0.0021 & 0.000 \\ -0.0290 & 0.0638 & -0.0320 & 0.0039 \\ 0.1169 & -0.2120 & 0.0605 & 0.2589 \end{bmatrix} \begin{bmatrix} \theta_s^3 \\ \theta_s^2 \\ \theta_s \\ 1 \end{bmatrix} \\ y_z &= \begin{bmatrix} T^2 & T & 1 \end{bmatrix} \begin{bmatrix} 0.0028 & -0.0061 & 0.0032 & 0.000 \\ -0.0421 & 0.0897 & -0.0415 & 0.0052 \\ 0.1535 & -0.2676 & 0.0667 & 0.2669 \end{bmatrix} \begin{bmatrix} \theta_s^3 \\ \theta_s^2 \\ \theta_s \\ 1 \end{bmatrix} \end{aligned} \tag{A.4}$$

Note that in the *Lalonde-Matthews* model [19], they removed normalization by zenith luminance in Eq. A.1, but instead fit its color weights directly to the observed sky data. Still they use the same formular in [77] to determine the five coefficients with respect to the turbidity T. The coefficients for the Y, x and y distribution functions are:

$$\begin{bmatrix} A_Y \\ B_Y \\ C_Y \\ D_Y \\ E_Y \end{bmatrix} = \begin{bmatrix} 0.1787 & -1.4630 \\ -0.3554 & 0.4275 \\ -0.0227 & 5.3251 \\ 0.1206 & -2.5771 \\ -0.0670 & 0.3703 \end{bmatrix} \begin{bmatrix} T \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} A_x \\ B_x \\ C_x \\ D_x \\ E_x \end{bmatrix} = \begin{bmatrix} -0.0193 & -0.2592 \\ -0.0665 & 0.0008 \\ -0.0004 & 0.2125 \\ -0.0641 & -0.8989 \\ -0.0033 & 0.0452 \end{bmatrix} \begin{bmatrix} T \\ 1 \end{bmatrix} \tag{A.5}$$

$$\begin{bmatrix} A_y \\ B_y \\ C_y \\ D_y \\ E_y \end{bmatrix} = \begin{bmatrix} -0.0167 & -0.2608 \\ -0.0950 & 0.0092 \\ -0.0079 & 0.2102 \\ -0.0441 & -1.6537 \\ -0.0109 & 0.0529 \end{bmatrix} \begin{bmatrix} T \\ 1 \end{bmatrix}$$

Therefore, the f_P is $\mathcal{F}(\theta, \gamma)$ in Eq. A.2 where its coefficients are determined with the turbidity T and the corresponding color component.