Towards Scalable Multi-View Reconstruction of Poses, Geometry, and Materials

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von

Carolin Marina Schmitt

aus Hanau

Tübingen
2023

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	04.12.2023
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Andreas Geiger
2. Berichterstatter:	Prof. Dr. Hendrik Lensch

Abstract

Accurately reconstructing the geometry and materials of indoor scenes from 2D images is a challenging research problem in Computer Vision and Computer Graphics. It is particularly difficult for input data captured by mobile acquisition systems of larger, multi-object scenes. The main challenges arise from the strong correlation between geometry and material parameters, the scale of the scene, and the sparse samples obtained from handheld data. In this dissertation, we address all these problems by presenting a novel method for joint recovery of object geometry, material reflectance, and camera pose of 3D scenes that exceed object-scale. The input is high-resolution RGB-D images captured by a mobile, handheld sensor system with active illumination by point lights.

Recovering scene parameters given RGB-D images is ill-posed due to the strong link between geometry and material parameters. The image formation process renders the appearance of a scene in 2D by estimating the light hitting the camera. In detail, it models the light that gets reflected from the scene surfaces based on the surface orientation, geometrical scene configuration, and material properties. Here, different combinations of geometry and material parameter estimates can result in the same appearance on a pixel level. To recover these scene parameters from appearance, the image formation needs to be reversed – a highly under-constraint problem. While previous works estimate geometry and material properties in alternation, these correlated entities are best optimized jointly. Therefore, we formulate the problem using a single objective function that can be minimized jointly using off-the-shelf gradient-based solvers, resulting in cleanly separated parameters.

Next, accurately reconstructing a scene that exceeds object scale requires a scalable scene representation, i.e., one that allows for the optimization of large numbers of input views and many parameters while still being computationally feasible and memory-efficient. To this end, we introduce a scene representation based on a set of local 2.5D keyframes and a distributed optimization algorithm; together, these enable accurate scene reconstructions with a memory footprint that does not scale with the scene size. Additionally, our novel multi-view consistency regularizer effectively synchronizes neighboring keyframes, allowing seamless integration into a globally consistent 3D model.

Finally, sparse samples obtained from handheld capture systems contain insufficient information to estimate material parameters accurately. Therefore, prior knowledge in the form of carefully designed regularizers must be added to the optimization objective. We present a novel smoothness term that effectively propagates material information over the scene surface while preserving clean material boundaries. We thus achieve accurate parameter estimates and realistic appearance reconstruction from sparse observations, even for challenging materials like glossy surfaces and specular highlights.

Backed up by thorough ablations and experiments, we believe this work is a valuable step towards large-scale, indoor 3D reconstruction of poses, geometry, and materials.

Kurzfassung

Sowohl die Geometrie als auch die Materialien von Innenräumen anhand von 2D Bildern exakt zu rekonstruieren, ist eine der Forschungsfragen im maschinellen Sehen und in der Computergrafik. Besonders größere Szenen mit mehreren Objekten und Daten von tragbaren Aufnahmegeräten stellen Herausforderungen dar. Die Hauptprobleme hierbei sind die starke Korrelation von Geometrie- und Materialparametern, die Komplexität und Größe der Szene die rekonstruiert werden soll und die spärlichen Datenpunkte von mobilen Aufnahmegeräten. In dieser Dissertation befassen wir uns mit all diesen Problemen, indem wir eine neue Methode zur gleichzeitigen Rekonstruktion von Objektgeometrie, Materialienreflektionen und Kameraposen für Szenen mit mehreren Objekten vorstellen. Der Input besteht aus hoch auflösenden RGB-D Bildern, die mit einem portablen, handgeführten Sensorsystem aufgenommen werden, wobei die Szene aktiv mit Hilfe von Punktlichtquellen beleuchtet wird.

Szenenparameter aus RGB-D Bildern zu rekonstruieren ist auf Grund der starken Abhängigkeit von Geometrie- und Material-Parametern nicht eindeutig möglich. Der bildgebende Prozess berechnet das Erscheinungsbild einer Szene in 2D, indem er die Lichtstrahlen, die die Kamera treffen, schätzt. Dabei wird das Licht modelliert, dass die Oberflächen der Szene reflektieren, in Abhängigkeit von den Oberflächenausrichtung, den Materialeigenschaften und der geometrischen Konfiguration der Szene. Diese Berechnung des Bildes kann, vor allem auf Pixel-Ebene, für verschiedene Kombinationen von Geometrie- und Material-Parametern zu dem gleichen Aussehen führen. Dadurch ist der umgekehrte Prozess, welcher die Parameter anhand von Erscheinungsbildern rekonstruiert, stark unterbestimmt. Während frühere Arbeiten Geometrie und Materialien alternierend rekonstruierten, sind wir überzeugt, dass mit gleichzeitiger Optimierung diese starke Korrelation besser abgebildet wird. Daher formulieren wir das Optimierungsproblem mit einer einzigen Zielfunktion, welche mittels existierender, Gradienten-basierter Algorithmen minimiert werden kann. Unsere Ergebnisse belegen, dass gleichzeitige, im Gegensatz zu alternierender Optimierung bessere Ergebnisse und klarer separierte Parameter Schätzungen liefert.

Des Weiteren erfordert die Rekonstruktion von Szenen mit mehreren Objekten eine skalierbare Szenenrepräsentation. Insbesondere muss diese Repräsentation die Optimierung einer großen Zahl von Input Bildern und vielen Parametern ermöglichen, ohne Rechenleistung oder Speicherkapazitäten zu überschreiten. Unsere Lösung besteht aus einer Kollektion lokaler 2.5D Keyframes und einem verteilten Optimierungsalgorithmus. Zusammen ermöglichen diese akkuraten Rekonstruktionen, mit einem Speicherbedarf der nicht mit der Szenengröße skaliert. Hierbei synchronisiert unsere neue Multi-View-Konsistenz-Regularisierung effektiv benachbarte Keyframes und ermöglicht so die nahtlose Integration in ein global konsistentes 3D-Modell.

Schließlich enthalten die spärlichen Datenpunkte mobiler Aufnahmen nur ungenügende

Informationen zur genauen Schätzung von Materialparametern. Daher muss Vorwissen, vorzugsweise in Form von sorgfältiger Regularisierung der Zielfunktion, integriert werden. Wir stellen einen neuen Glättungsterm vor, welcher die Materialinformation der Datenpunkte effektiv über die Szenenoberfläche propagiert und dabei Materialiengrenzen berücksichtigt. Auf diese Weise erreichen wir exakte Parameterschätzungen trotz spärlicher Datenpunkte, und rekonstruieren realistische Bilder, selbst für herausfordernde Materialien, wie glänzende Oberflächen und spekulare Glanzlichter.

Gestützt von gründlichen Experimenten, sind wir der Meinung, dass diese Arbeit einen wertvollen Schritt hin zu einfachen 3D Rekonstruktionen von Kameraposen, Geometrien und Materialien komplexer Indoorszenen darstellt.

Publications

Publications [Sch+23] and [Sch+20] are covered in this dissertation:

- **C. Schmitt**, B. Antić, A. Neculai, J. H. Lee, and A. Geiger. "Towards Scalable Multi-View Reconstruction of Geometry and Materials". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023). DOI: 10.1109/TPAMI.2023.3314348
  Project page: https://sites.google.com/view/material-fusion/

- **C. Schmitt**, S. Donné, G. Riegler, V. Koltun, and A. Geiger. "On Joint Estimation of Pose, Geometry and svBRDF from a Handheld Scanner". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020. DOI: 10.1109/CVPR42600.2020.00355
  Project page: https://avg.is.mpg.de/publications/schmitt2020cvpr
  Code repository: https://github.com/autonomousvision/handheld_svbrdf_geometry/

Other publications during the Ph.D. are [Tos+21] and [Pas+18]. Both are not covered in this dissertation:

- F. Tosi, Y. Liao, **C. Schmitt**, and A. Geiger. "SMD-Nets: Stereo Mixture Density Networks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021. DOI: 10.1109/CVPR46437.2021.00883

- D. Paschalidou, A. O. Ulusoy, **C. Schmitt**, L. van Gool, and A. Geiger. "RayNet: Learning Volumetric 3D Reconstruction with Ray Potentials". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018. DOI: 10.1109/CVPR.2018.00410

I declare that this thesis has been created by me based on my own original research mentioned above. All projects result from collaborative efforts, and all authors provided critical revisions to the manuscripts. My advisor Prof. Dr. Andreas Geiger, was involved in all projects. He contributed guidance, ideas, text, and illustrative figures to the publications. Additionally, he built our hardware, the multi-sensor capture rig.

Dr. Vladlen Koltun was involved in our work on [Sch+20] and accompanied the project as senior scientific advisor, contributing ideas and guidance. The whole project that led to the publication [Sch+20] was a collaboration with Dr. Simon Donné, with equal contribution of him and me. Both of us contributed to all parts of the project: Data generation, ideas, code, experiments, text, and figures. Further, Dr. Gernot Riegler contributed ideas and some code.

During our work on [Sch+23], Prof. Dr. Joo Ho Lee contributed ideas, text, and some code. My co-author Božidar Antić contributed to the experiments, text, and code, and

Andrei Neculai contributed code and text of Sec. 5.5 to the publication. I played a critical role in developing this project and driving it forward. My contributions were significant in all aspects: Ideas, code, experiments, text, and figures. I was also in charge of all data generation, and both Joo Ho Lee and Božidar Antić assisted me in data capturing and running all data pre-processing.

In this dissertation, [Sch+20] is presented in Chap. 4 and [Sch+23] in Sec. 5. Further, parts of both works are included in Chap. 3.

Acknowledgments

I would like to express my sincere gratitude to my advisor Andreas Geiger, for his unwavering support, patience, and motivation throughout all these years. He consistently believed in our project and was a constant source of ideas on what to try next. I am highly thankful for his guidance and the chance to work in such a vibrant environment full of scientific curiosity, which enabled me to grow as both a person and a researcher.

My honest thanks also goes out to Hendrik Lensch and Georg Martius, who, along with Andreas Geiger, formed my *Thesis Advisory Committee (TAC)*. Thank you for your valuable advice, keen interest, for asking the right questions at the right time, and overall, your academic guidance! I would also like to extend a special thanks to Hendrik Lensch, my second supervisor, and his team for the engaging discussions and the generous offer to share their laboratory with us for our experiments.

I am immensely grateful for my co-authors, colleagues, and collaborators, without whom the projects would not have been possible. In alphabetical order: Andrei Neculai, Božidar Antić, Gernot Riegler, Joo Ho Lee, Simon Donné, and Vladlen Koltun. It was an honor to work with you!

Another big thanks to the whole *Autonomous Vision Group (AVG)* for the friendly atmosphere, opportunities for discussion, and passion for research.

Further, I would like to thank the *International Max-Planck Research School for Intelligent Systems (IMPRS-IS)*, an inspiring community of great people.

And a serious applause goes to service groups and all staff members of the *Max-Planck Institute for Intelligent Systems (MPI-IS)*! Especially, but not limited to, the *Optics and Sensing Laboratory* and the *Tübingen IT*. Thank you for providing and maintaining this excellent research environment and for almost instant help in all kinds of matters.

Overall, I am immensely happy for all the wonderful people I had the pleasure to meet in Tübingen, particularly at the Max-Planck Institute – so much enthusiasm, friendliness, ideas, diversity, inspiration, happiness, and fun. We created incredible memories in everyday life and on work-related and non-work-related trips. The friendships I found are invaluable!

Last, to my family: Endless love and happiness!

Contents

List of Figures

List of Tables

List of Algorithms

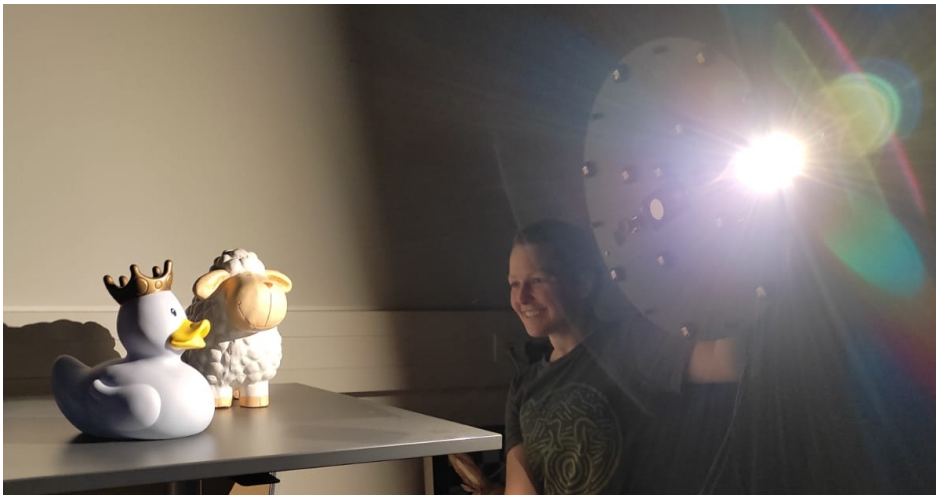
1 Introduction

Achieving photorealism, a point at which synthesized images reproduce the appearance of the real-world like a photograph is one of the big goals of various Computer Vision and Computer Graphics problems. The tasks range from 3D reconstruction over relighting, view synthesis, generative modeling, and intrinsic images to rendering. For most of them, photorealism necessitates reasoning about light-dependent effects; for that, some understanding of the scene and its content is very beneficial. We aim to use 3D reconstruction to infer scene information from RGB-D images, allowing for synthesizing new images from arbitrary viewpoints and under novel lighting.

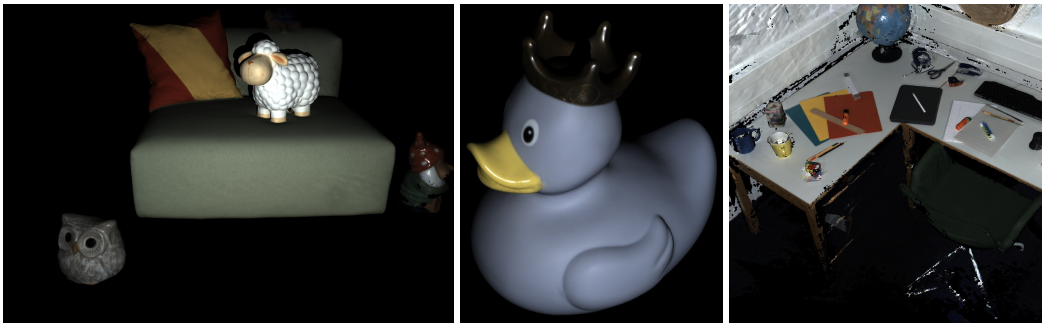
Ideally, we could disentangle geometry, materials, poses, and illumination for full controllability of the synthesized image. However, this is very challenging even for known illumination due to the interlaced and complex correlation of geometric and photometric entities. Most accurate results can be obtained from sophisticated light stages and in laboratory environments. But for most applications, like training embodied agents or telepresence, this is not practical: Typically, the scenes we are interested in exceed object-scale and cannot be put into light stages. This has two implications: First, reconstructing spatially larger scenes at a resolution that preserves detail requires a scene representation for which the computation memory does not scale with the scene size. Second, the only data easily available is captured by portable (likely handheld) camera systems. Thus, it does not contain calibrated camera poses nor dense coverage of, e.g., reflectance, both of which are typical for light stages. Instead, handheld data is captured from arbitrary viewpoints and only covers the scene sparsely. This is especially problematic for recovering view-dependent material properties like specular highlights.

In this dissertation, we present solutions to all the above problems: Our methods provide accurate geometry and material reconstructions from data captured by practical, mobile camera systems and scenes that exceed object scale, see Fig. 1.1. We further discuss the practical challenges of photorealism for 3D reconstructions.

We start by developing a reconstruction algorithm that jointly optimizes geometry and material parameters. These entities are heavily correlated: A good model of light transport and material-dependent reflections allows for recovering geometric detail using shading cues. An accurate shape model, in turn, facilitates the estimation of material properties. To solve this chicken-and-egg problem, previous works decompose it into smaller problems using multiple decoupled objectives and alternate the optimization of geometry and material parameters. In contrast, we found that jointly optimizing the correlated components is beneficial for accurate geometry and material reconstruction and fosters clean separation of parameters. Therefore, we formulate one optimization problem with a single objective function that can be optimized using off-the-shelf gradient-based solvers.



(a) *Data Capture: Handheld Sensor Rig with Active Illumination.*



(b) *Renderings of our Reconstruction Results for new Views and unseen Lighting.*

Figure 1.1: 3D Reconstruction from Handheld Data (3D). *Given data captured with a mobile, handheld camera system (a), reconstruct accurate geometry and material parameters that allow for synthesizing images at unknown viewpoints under novel lighting (b).*

A second problem is the optimization itself. Reversing the image formation process and estimating geometry and materials from images of multi-object scenes is intrinsically non-convex and ambiguous. A pixel appearing darker in an image can be caused by either a shifted surface normal, a greater distance to the camera, a darker material color, or a less glossy material. We tackle this ambiguity with a block-based optimization scheme and a soft regularizer, which promote global consistency by propagating multi-view information across neighboring blocks. During optimization, this builds a connected graph over all surface points of the scene. Combined with careful calibration and parameter initialization, this resolves many ambiguities effectively and causes a more stable optimization behavior.

Next, we propose a regularizer that promotes the accurate reconstruction of parameter maps for glossy materials given sparse samples from a handheld capture system. Observations captured by hand from arbitrary poses naturally exhibit more motion blur and less

dense coverage than images taken by, e.g., a robotic arm. This complicates the material estimation problem: Any physically-based reflectance function for a specific material is high-dimensional since it depends at least on both incoming and outgoing light directions and the surface orientation. Even when capturing a scene very thoroughly, the acquired samples of the reflectance function are sparse. Especially for glossy materials, these samples are insufficient for a purely data-driven recovery of this high-dimensional function. Therefore, strong priors or effective regularizers are necessary for accurate reconstructions.

Some prior works reduce the dimensionality of the reflectance function by assuming knowledge of the surface orientation (e.g., limit to flat surfaces), the materials present in the scene, or by restricting the captures to co-located setups of light source and camera. Instead, we present a regularization-based approach that enforces the propagation of reflectance parameters across pixels by assuming that nearby pixels with similar diffuse behavior also exhibit similar specular behavior. We demonstrate that this effectively links surface areas of the same material, leading to clean and detailed material parameter estimates despite sparse observation samples.

Finally, we present a distributed optimization scheme over non-global scene representations. It features a constant optimization memory footprint independent of the scene size while enabling accurate global integration. Unfortunately, reconstructing larger multi-object scenes captured from many viewpoints at high resolution (e.g., 4K) becomes intractable quickly. In order to process such scenes, a scalable scene representation is inevitable. Therefore, we propose using local 2.5D scene representations and a distributed, decentralized optimization scheme that encourages global consistency between the representations. We show that, despite overlapping fields of view, regularizing multi-view consistency is crucial to attaining globally accurate reconstructions without visual artifacts. By optimizing in 2.5D, the computation memory of the proposed model is independent of the scene size and allows for reconstructing geometry and materials at larger scales.

This dissertation is organized as follows: We start by summarizing background knowledge on photo-realistic rendering and scene representations in Chap. 2. This is followed by a review of the most relevant related work on differentiable rendering as well as geometry and material estimation, both separately, jointly, and at scale, in Chap. 3. Chap. 4 presents our work '*On Joint Estimation of Pose, Geometry and svBRDF from a Handheld Scanner*', published at CVPR 2020, [Sch+20]. Our follow-up work, '*Towards Scalable Multi-View Reconstruction of Geometry and Materials*', is accepted in TPAMI [Sch+23] and presented here in Chap. 5. Following the publications, we identify the main challenges for accurate geometry and reflectance estimation for indoor scenes and add a thorough discussion as well as an outlook in Chap. 6.

2 Background

In this chapter, we introduce relevant background knowledge on photo-realistic rendering in Sec. 2.1 and 3D scene representations in Sec. 2.2.

2.1 Photo-Realistic Rendering

Photo-realistic rendering is the process of synthesizing 2D images from virtual 3D scenes that are visually indistinguishable from reality captured in 2D photos. Real images are created by light hitting a scene and being reflected off surfaces and volumetric media possibly several times, before entering camera optics and being detected by the sensor. One way to create images that mirror real images is to simulate this process by precisely modeling all physical objects and phenomena. It includes a detailed 3D scene, lights, a camera, its optics, and direct and indirect light transport with all visibility constraints as well as surface and volume scattering events. This enables editable, interpretable, and physically based renderings. Another approach is to leverage generative neural networks' simplicity and great representation power and learn a photo-realistic rendering engine from data. With sufficient and informative data, these renders can recover, e.g., scattering phenomena that are challenging to model analytically. While these first classical methods quickly become very complex, it is hard to guarantee physical plausibility for the latter neural methods. As we target to reconstruct physical parameters, this dissertation focuses on classical rendering methods.

Photorealistic rendering is a very active field of research in Computer Graphics. Applications include, e.g., game and movie production, augmented and virtual reality applications, architecture, simulators, design visualization, and many more. Modern rendering engines produce astonishing results which made them a powerful tool and central component for many Computer Vision and Machine Learning tasks. Especially for 3D reconstructions, the availability of accurate 2D renderings enables quantitative evaluation based on pixel differences between synthetic and real images. This *photometric loss* is the primary metric for reconstruction quality we use in this dissertation. It serves as our measure of *photorealism* and '*indistinguishability from reality*', and we also include a discussion in Sec. 6.1.3.

This section introduces photo-realistic rendering for the task of 3D reconstruction. We start with the physical background by defining relevant radiometric entities in Sec. 2.1.1. We then derive a mathematical formulation for light reflection at surfaces in Sec. 2.1.2 and follow up by introducing existing models to approximate reflectance: Sec. 2.1.3 is on data-driven, and Sec. 2.1.4 on parametric reflectance models. We conclude this section by presenting the rendering equation and estimation techniques in Sec. 2.1.5.

2.1.1 Radiometric Entities

Radiometry is the study of the propagation of electromagnetic radiation in an environment and its measurement. It provides mathematical tools to describe the physical phenomena of light propagation and reflection. There are many possible interactions of light with materials; here, we focus on the ones relevant to scans of steady indoor scenes. This means we assume linear light (the output beam is assumed to be a linear function of the input beam, not valid for, e.g., lasers) in steady-state or equilibrium (constant over time) and ignore the polarization of the electromagnetic field as well as effects of fluorescence or phosphorescence (which belong to the field of quantum optics). In this section, we describe light as photons since all radiometric quantities can be well defined in terms of photon events - they are ways of measuring photons. This section, particularly the notations and definitions, is based on [PJH16; Hug+13; Vea98].

Light Energy: The following radiometric entities can be evolved from the knowledge that light photons carry energy. We denote the photon energy Q; it is given by

$$Q(\lambda) = \frac{hc}{\lambda} \tag{2.1}$$

where c is the speed of light, h Planck's constant and λ the wavelength.

In this dissertation, we are mainly concerned with color images from RGB cameras. Therefore, we will adapt the RGB color model in Sec. 2.1.4 and represent the light spectrum by integration over three wavelength intervals. Thus, while the theory presented here is general to all light models, we omit the parameter λ from the notation of $Q(\lambda)$ in the following.

Flux: In rendering, we are interested in the total emission of light sources, which is the emitted light energy per time interval. The flux Φ (also radiant flux or power) is the total energy passing through a region of space in unit time $\Phi = \frac{Q}{t}$. It can be thought of as photon density. In differential form, it is defined as:

$$\Phi(t) = \lim_{\Delta t \to 0} \frac{\Delta Q}{\Delta t} = \frac{\mathrm{d}Q}{\mathrm{d}t} \tag{2.2}$$

Conversely, integrating flux Φ over time yields the total energy Q:

$$Q = \int_t \Phi(t)\ \mathrm{d}t \tag{2.3}$$

The unit of flux is joules per second or watts $\frac{J}{s} = W$. Note that since we assume systems in equilibrium, Φ is constant over time, and the parameter t can be omitted from the notation.

Irradiance: Flux Φ can only be measured over an area A. Therefore, the average power density per unit surface area is defined as irradiance $E = \frac{\Phi}{A}$. Intuitively, it is the area density of flux. Note that the term irradiance E usually refers to light that is arriving at a surface; light leaving a surface is called radiant exitance or radiosity M for distinction. In differential

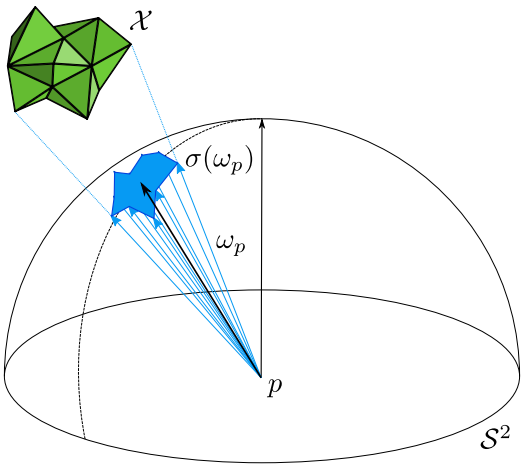


Figure 2.1: Solid Angle. *For a point p, the solid angle subtended by an object $\mathcal{X}$ is equal to the area of the object's projection onto the unit sphere centered at p.*

form, both are defined as:

$$E(A) = \lim_{\Delta A \to 0} \frac{\Delta \Phi}{\Delta A} = \frac{\mathrm{d}\Phi}{\mathrm{d}A} \tag{2.4}$$

or in integral form

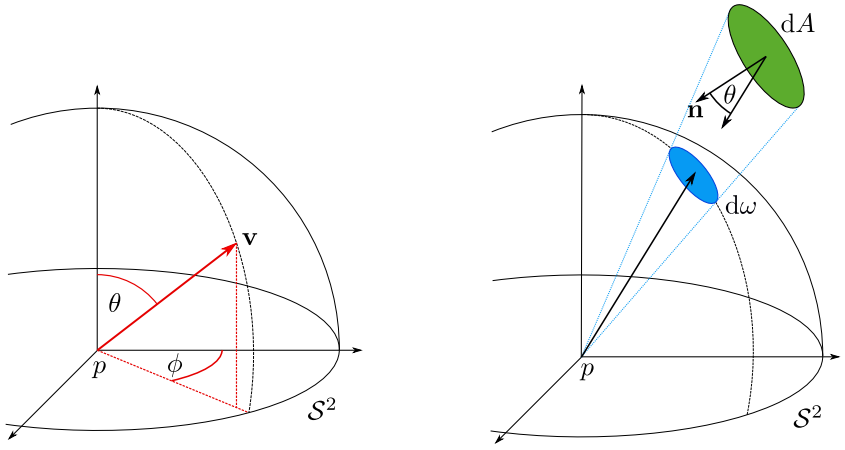
$$\Phi = \int_A E(A)\ \mathrm{d}A \tag{2.5}$$

Example: A point light is a point in space that emits the same amount of energy in all directions. Assume such a point light $l \in \mathbb{R}^3$ and another point $p \in \mathbb{R}^3$ at a distance $r \in \mathbb{R}$ to the light. Then the flux or power Φ is distributed over the area of a sphere at point l with radius r and the irradiance at point p is given by $E = \frac{\Phi}{4\pi r^2}$. Thus, the light received at point p scales quadratically with the distance.

Solid Angle: A 2D angle α can be measured by the length of the arc on the unit circle that the angle, placed in the center of the circle, intercepts. The unit of this representation for angles is radians (*rad*). The solid angle ω is the generalization of a 2D angle to 3D. It is measured by the area on the unit sphere surface that the solid angle, placed in the center of the sphere, intercepts. Its unit is steradians (*sr*). For example, a hemisphere (with arbitrary radius) subtends a solid angle of $2\pi\,sr$.

Formally, let $\sigma : \mathcal{S}^2 \to \mathbb{R}$ denote the common surface area measure on $\mathcal{S}^2$ and define a set of directions $\omega \subset \mathcal{S}^2$, then the solid angle occupied by ω is $\sigma(\omega)$. Further, the solid angle at a point p subtended by an object or surface $\mathcal{X}$ is given by the measure of the set of directions ω_p obtained by projecting $\mathcal{X}$ onto the unit sphere centered at p, see Fig. 2.1. To be consistent with other sources, we will abuse notation slightly and use symbol ω to denote both a set of directions on the unit sphere $\omega \subset \mathcal{S}^2$ and their solid angle $\sigma(\omega) \in \mathbb{R}$ (similar to symbol A which usually stands for either a surface $A \subset \mathbb{R}^2$ or the surface area $A \in \mathbb{R}$).

Two properties of solid angles are important for integration: First, we can transform an integral over a solid angle to an integral over spherical angles using the following: Let



(a) *Spherical Angles.* ***(b)*** *Solid Angle Subtended by a Surface.*

Figure 2.2: Properties of Solid Angles. *A solid angle can be expressed in Euclidean coordinates or spherical coordinates (a). Also, the differential solid angle* $\mathrm{d}\omega$ *at* p *is proportional to the differential area* $\mathrm{d}A$ *of the surface subtending it (b).*

$\mathbf{v} \in \mathcal{S}^2$ be a unit vector and (θ, ϕ) its polar and azimuthal angles, see Fig. 2.2a, then the differential solid angle $\mathrm{d}\omega$ around $\mathbf{v}$ corresponds to

$$\mathrm{d}\omega = \sin\theta \, \mathrm{d}\theta \, \mathrm{d}\phi \tag{2.6}$$

Second, we can also transform integrals over directions into integrals over the area: Assume a small surface patch S with normal $\mathbf{n}$ and infinitesimal surface area $\mathrm{d}A$ is seen from point p at a distance r. Let θ denote the angle between normal $\mathbf{n}$ and the vector from $\mathrm{d}A$ to p, see Fig. 2.2b. Then the differential solid angle $\mathrm{d}\omega$ subtended by surface S at p is related to the differential area $\mathrm{d}A$ by

$$\mathrm{d}\omega = \frac{\mathrm{d}A \cos\theta}{r^2} \tag{2.7}$$

Intensity: For an infinitesimal light source or a perfect point light, we can now define the angular density of power: the intensity. In contrast to flux and irradiance, intensity is a directional quantity. It is given as

$$I(\omega) = \lim_{\Delta\omega \to 0} \frac{\Delta\Phi}{\Delta\omega} = \frac{\mathrm{d}\Phi}{\mathrm{d}\omega} \tag{2.8}$$

And integrating the intensity over all directions gives the flux

$$\Phi = \int_{\mathcal{S}^2_+} I(\omega) \, \mathrm{d}\omega \tag{2.9}$$

The intensity unit is watts per steradian $\frac{W}{sr}$. For the full sphere of directions, the emitted intensity is $I = \frac{\Phi}{4\pi}$. When considering the observed intensity at a surface, Lambert stated

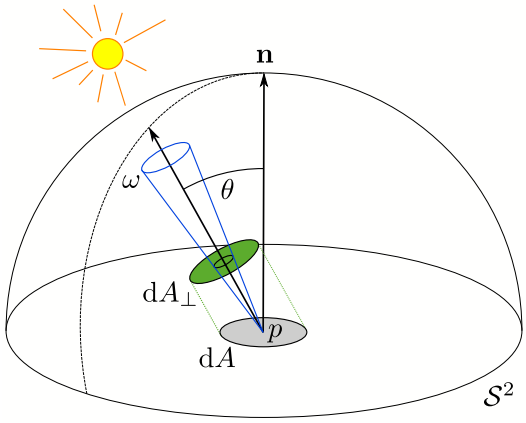


Figure 2.3: Incident Light on a Surface. *The light intensity incident on a surface with differential area* dA *depends on the angle* θ *between the incident direction* ω *and the surface normal* **n**. *For calculation, we project the surface patch onto a virtual plane perpendicular to the light direction, shown in green.*

what is now known as **Lambert's cosine law** [Lam60]: '*For a light source with intensity* I *that illuminates a surface, the intensity* $I(\theta)$ *observed at the surface is proportional to the cosine of the angle* θ *between the surface normal and the direction of the incident light.*'

$$I(\theta) = I\cos\theta \tag{2.10}$$

Likewise, for a patch on the surface with infinitesimal area dA, illuminated under the angle θ, we denote d$A_\perp$ the area of its projection onto a virtual plane perpendicular to the direction of the incident light, see Fig. 2.3. Then, it can be shown that:

$$\mathrm{d}A = \frac{\mathrm{d}A_\perp}{\cos\theta} \tag{2.11}$$

Radiance: The last quantity we define is radiance. While irradiance captures the total flux or light power arriving at a surface, radiance captures the directional distribution of the power by relating it to the solid angle of the incident direction. For a point p and a set of directions of incident light ω, denote $A_\perp$ the surface area perpendicular to ω and define

$$L(p,\omega) = \lim_{\Delta\omega\to 0}\frac{\Delta E(A_\perp(p))}{\Delta\omega} = \frac{\mathrm{d}E(A_\perp(p))}{\mathrm{d}\omega} \tag{2.12}$$

or equivalently

$$E(A_\perp(p)) = \int_{\mathcal{S}^2_+} L(p,\omega)\,\mathrm{d}\omega \tag{2.13}$$

One usually differs between incident radiance $L_{\text{in}}(p,\omega)$ and exitant or outgoing radiance $L_{\text{out}}(p,\omega)$. In both cases, we define ω as the direction vector that points away from the surface at point p. Note: This is inconsistent in literature, and notation varies slightly in

Radiometric	Unit	Photometric	Unit
Radiant energy	joule (Q)	Luminous energy	talbot (T)
Radiant flux	watt (W)	Luminous flux	lumen (lm)
Intensity	$\frac{W}{sr}$	Luminous intensity	candela $\left(cd = \frac{lm}{sr}\right)$
Irradiance	$\frac{W}{m^2}$	Illuminance	lux $\left(lx = \frac{lm}{m^2}\right)$
Radiance	$\frac{W}{m^2 sr}$	Luminance	nit $\left(nit = \frac{lm}{m^2 sr} = \frac{cd}{m^2}\right)$

Table 2.1: Radiometric Measurements and their Photometric Analogs. *From [PJH16].*

other sources. An essential property of radiance is that it is constant along rays through empty space. Therefore, $L_{\text{out}}(p,\omega) = L_{\text{in}}(p,-\omega)\ \forall\omega$ for a point p in free space.

Radiance is the quantity rendering is most concerned with; we see why in the next section. Before, consider once more Fig. 2.3: Light is arriving at an infinitesimal surface area $\mathrm{d}A(p)$ with normal $\mathbf{n}_p$ at surface point p under incident directions $\mathrm{d}\omega$. As before, denote $\mathrm{d}A_\perp(p)$ the projected differential area of $\mathrm{d}A(p)$ onto the virtual surface perpendicular to $\mathrm{d}\omega$. Then, with the previous definitions, we can express radiance as:

$$L(p,\omega) \overset{2.12}{=} \frac{\mathrm{d}E(A_\perp(p))}{\mathrm{d}\omega} \overset{2.4}{=} \frac{\mathrm{d}^2\Phi(p)}{\mathrm{d}\omega\,\mathrm{d}A_\perp(p)} \overset{2.11}{=} \frac{\mathrm{d}^2\Phi(p)}{\mathrm{d}\omega(\omega\cdot\mathbf{n}_p)\,\mathrm{d}A(p)} \overset{2.4}{=} \frac{\mathrm{d}E(A(p))}{\mathrm{d}\omega(\omega\cdot\mathbf{n}_p)} \tag{2.14}$$

Excursion: Radiometry and Photometry

Unlike the radiometric units in this section, photometric units like lumen, candela, and lux are commonly known and used to measure visible light. In fact, these units have a direct connection: Photometry describes the same quantities of electromagnetic radiation as radiometry but in terms of perception by the human visual system. Let $V(\lambda)$ denote the spectral response curve, i.e., the human eye's sensitivity to different wavelengths. Then, e.g., luminance Y with unit candelas per meter squared $\frac{cd}{m^2}$ is related to spectral radiance $L(\lambda)$ by

$$Y = \int_\lambda L(\lambda)V(\lambda)d\lambda \tag{2.15}$$

Likewise, each radiometric quantity is related to a corresponding photometric quantity. A summary can be found in Tab. 2.1.

2.1.2 Surface Reflection

When light hits a surface, various kinds of scattering events happen. Depending on the wavelength spectrum, polarization, and intensity of the light and the micro-structure and molecules of a surface, an incoming light ray might get reflected, transmitted, refracted, retro-reflected, absorbed, and scattered again within the surface. Since we are mainly concerned with light-material interactions in indoor environments without humans or animals, we focus on opaque and non-translucent materials. Additionally, we do not use polarized lights in our experiments. Therefore, polarization and subsurface light transport are less relevant and

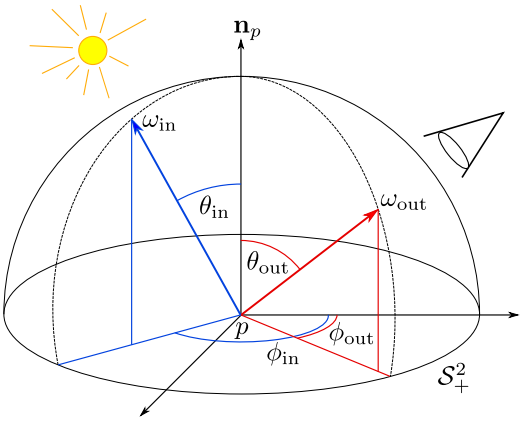


Figure 2.4: The BRDF. *The BRDF f_p models the fraction of light that is reflected from incoming light direction ω_{in} to outgoing light direction ω_{out} given the surface normal $\mathbf{n}$ at point p.*

not considered in this dissertation. Instead, we concentrate on the directional and spectral distribution of reflected light. It includes, e.g., why a blue couch can look blue from all directions, why a mug shows very bright highlights that vary with the view direction, and why a mirror reflects its surroundings, completely changing its color and appearance with the viewpoint.

In the previous section, we discussed the propagation of light incident on a surface. In this section, we trace the light path further and look at the effects after light has hit a surface. We set the ground by deriving a mathematical formulation for light reflection at a surface next.

Bidirectional Reflectance Distribution Function (BRDF): Consider illuminating a surface at point p with normal $\mathbf{n}_p$ and the incident light arrives from an infinitesimal cone of directions that occupies a differential solid angle $\mathrm{d}\omega_{\text{in}}$, see Fig. 2.4. Then recall Eq. (2.14): The radiance along ω_{in} hitting p relates to the total flux incident at the surface following

$$\mathrm{d}E(p,\omega_{\text{in}}) = L_{\text{in}}(p,\omega_{\text{in}})(\mathbf{n}_p \cdot \omega_{\text{in}})\,\mathrm{d}\omega_{\text{in}} \tag{2.16}$$

Our assumption of a linear optical system implies that this differential irradiance striking the surface causes a differential outgoing radiance leaving the surface and that the relationship is linear. The resulting proportionality $\mathrm{d}L_{\text{out}}(p,\omega_{\text{out}}) \propto \mathrm{d}E(p,\omega_{\text{in}})$ gives rise to the definition of the *Bidirectional Reflectance Distribution Function*, in short *BRDF*:

$$\begin{aligned} \rho : \mathbb{R}^3 \times \mathcal{S}^2 \times \mathcal{S}^2 &\rightarrow \mathbb{R}^3 \\ \rho(p,\omega_{\text{in}},\omega_{\text{out}}) &= \frac{\mathrm{d}L_{\text{out}}(p,\omega_{\text{out}})}{\mathrm{d}E(p,\omega_{\text{in}})} \overset{2.16}{=} \frac{\mathrm{d}L_{\text{out}}(p,\omega_{\text{out}})}{L_{\text{in}}(p,\omega_{\text{in}})(\mathbf{n}_p \cdot \omega_{\text{in}})\,\mathrm{d}\omega_{\text{in}}} \end{aligned} \tag{2.17}$$

Intuitively, the BRDF describes the fraction of all incoming light reflected in a specific direction. The BRDF has two important properties: The first is *reciprocity* (also called

Helmholz reciprocity): The BRDF is symmetric in the light and view directions:

$$\rho(p, \omega_{\text{in}}, \omega_{\text{out}}) = \rho(p, \omega_{\text{out}}, \omega_{\text{in}}) \tag{2.18}$$

And the second is *energy conservation*: The total reflected light power must not exceed the total incident light power. Effectively, this restricts the BRDF to fulfill

$$\int_{\mathcal{S}^2_+} \rho(p, \omega_{\text{in}}, \omega_{\text{out}})(\mathbf{n}_p \cdot \omega_{\text{in}}) \, \mathrm{d}\omega_{\text{out}} \leq 1 \qquad \forall \omega_{\text{in}} \in \mathcal{S}^2_+ \tag{2.19}$$

with $\mathcal{S}^2_+$ denoting the upper unit hemisphere around normal $\mathbf{n}_p$. On a side-note: The function describing transmission (*Bidirectional Transmission Distribution Function (BTDF)*) is very similar to Eq. (2.17) for reflection, but the two properties of reciprocity and energy conservation are unique to reflection and do not always apply to transmission.

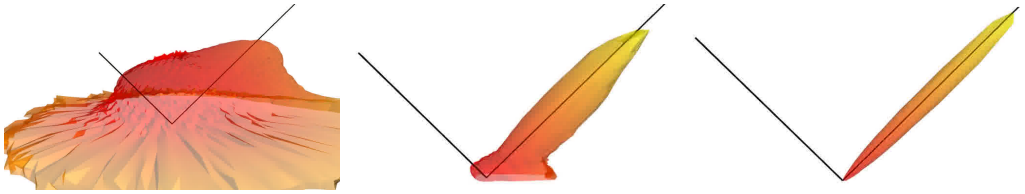
With the BRDF ρ defining the proportionality function in Eq. (2.17), we can now predict the outgoing radiance $L_{\text{out}}(p, \omega_{\text{out}})$ by integration over all directions in the upper unit hemisphere $\mathcal{S}^2_+ \subset \mathcal{S}^2$ around p, with the normal $\mathbf{n}_p$ pointing to its pole (also $\mathcal{H}^2(\mathbf{n}_p)$ in literature):

$$L_{\text{out}}(p, \omega_{\text{out}}) = \int_{\mathcal{S}^2_+} L_{\text{in}}(p, \omega_{\text{in}}) \, \rho(p, \omega_{\text{in}}, \omega_{\text{out}})(\mathbf{n}_p \cdot \omega_{\text{in}}) \, \mathrm{d}\omega_{\text{in}} \tag{2.20}$$

It is the basis for the rendering equation, which we introduce in Sec. 2.1.5.

Approximations of the BRDF: The BRDF for a given point p is a 4D function that describes the reflection behavior of light at surfaces. As such, it is a material property. For certain classes of materials or specific acquisition systems, additional properties and priors can simplify the function: For one, many real-world materials show **isotropic** reflectance, which means that the BRDF is symmetric around the surface normal $\mathbf{n}_p$. This reduces the BRDF to a 3-dimensional function with parameters $\theta_{\text{in}}, \theta_{\text{out}}$ and $\phi = \phi_{\text{in}} - \phi_{\text{out}}$. Anisotropic materials include, e.g., brushed metal and velvet. Then, a popular approach is to capture BRDF samples with **co-located camera and lighting**, e.g., using a phone and its flash. Here, a common simplification is to assume $\omega_{\text{in}} = \omega_{\text{out}}$ for the captured reflectance samples, resulting in a 2-dimensional BRDF. Last, some (indeed very few) materials are well approximated by a constant BRDF. This is called the **Lambertian** assumption because it follows Lambert's Law for ideal diffusely reflecting surfaces, Eq. (2.10). It is prevalent in early works on reflectance or geometry reconstruction due to its simplicity. But unfortunately, this assumption does not hold for the majority of real-world materials and only provides rough approximations.

Measured BRDFs: Measuring the BRDF of a material is very challenging and laborious. It requires a specialized and carefully calibrated acquisition device and many dense measurements to represent the BRDF as comprehensively as possible. An early and very influential, large-scale dataset of measured BRDFs was presented by Matusik et al. in 2003 [Mat+03], the *MERL database*. It contains dense measurements of about 130 real, isotropic materials using a specialized, custom-built measurement device. Based on the coordinate system



(a) Red Plastic *(b) Cherry* *(c) Color Changing Paint 2*

Figure 2.5: Measured BRDF. *Real measurements of three isotropic material samples from the MERL database [Mat+03]. Under an incident angle of $\theta_{in} = 45°$ (light coming from the left), we show the measured BRDF as a 3D plot: Each point θ_{out}, ϕ on the unit sphere is scaled with radius $r = \rho_{\theta_{in}}(\theta_{out}, \phi)$ and radial distances are color-coded. As expected, the highest values are observed under the perfect mirror reflection direction. Measurements show that the variance between BRDFs of different materials is large.*

proposed by Rusinkiewicz [Rus98], they divide space into approximately 1.5 million bins per color channel and material for capturing. The result is very high-dimensional, tabulated BRDF data; examples are shown in Fig. 2.5. Dupuy et al. [DJ18] captured another BRDF dataset: Using a motorized goniophotometer with a customized lighting setup, they measure spectral BRDFs of 32 isotropic and four anisotropic materials (including, e.g., the silk from a sari woven from two differently-colored yarns). Both use flat or spherical samples of homogeneous material targets.

Such datasets of BRDF measurements from sophisticated data acquisition systems are the only existing real ground truth for material and BRDF estimation. In the next section, we present models that efficiently approximate the BRDF and enable material estimation from less complete scans of objects and scenes composed of multiple materials.

Modeling the BRDF: For each surface point p, the BRDF is a 4-dimensional function, and until now, there is no '*go-to*' or '*standard*' model that is superior over the others. Each model has different advantages and weaknesses, is designed for different tasks or use cases, and defines a specific space of BRDFs that it can represent.

Generally, one differs between data-driven models that are based on BRDF measurements of material samples (presented below in Sec. 2.1.3) and parametric models that, often originating from physical or phenomenological observations, approximate the BRDF by a parametric function (presented in Sec. 2.1.4).

2.1.3 Data-Driven BRDF Models

Data-driven BRDF models typically use measured BRDF databases or dictionaries, as presented in Sec. 2.1.2. To compress high-dimensional BRDF measurements into a low-dimensional representation, there are multiple approaches: Matusik et al. [Mat+03] apply classical linear and non-linear dimensionality reduction methods, Lawrence et al. [Law+06] propose inverse shade trees to factor measured data into components, Nielsen et al. [NJR15] find a mapping of the BRDF space that allows recovering descriptive principal components, and Hu et al. [Hu+20] train an autoencoder to learn a non-linear latent embedding of

measured BRDF data. Similarly, Boss et al. [Bos+21a] and Zhang et al. [Zha+21c] present approaches that train *Multi-Layer Perceptrons (MLPs)* to predict the BRDF per object and pre-train these networks on the MERL database. Both compress their material model by restricting the latent space of their networks to two or three dimensions. Given any such low-dimensional embedding of the BRDF measurements, per-point reflectance is represented as a point in the embedding space.

A common alternative is to solve for a set of base BRDFs from the measured database, then estimate weights per-point/pixel or object, and linearly combine these base BRDFs. Here, [Li+20a; AZK08; ZWT13; Ren+11; Hui+17; HS15] pre-define the number of base BRDFs by hand, usually a low number between 1 and 10, while [GPG14] solve for it with an iterative clustering and fitting scheme. Note that the idea of a set of base BRDFs is equally used with parametric BRDF models, e.g., [HLZ10; WYT16; Len+03].

The big advantage of data-driven BRDF models is that they include effects that are difficult to model with parametric BRDF models, like anisotropy, metallicness, or retro-reflection. It requires balancing expressiveness and compression. A disadvantage of measured BRDF models for rendering is that they cannot be sampled explicitly, making it less effective and slower compared to parametric models. Further, data-based material models do not provide any intuitive understanding or options for easy manipulation: Materials can be blended, but how to render an appearance *more glossy* or *less yellow* is not straight-forward which makes them less attractive to, e.g., artists.

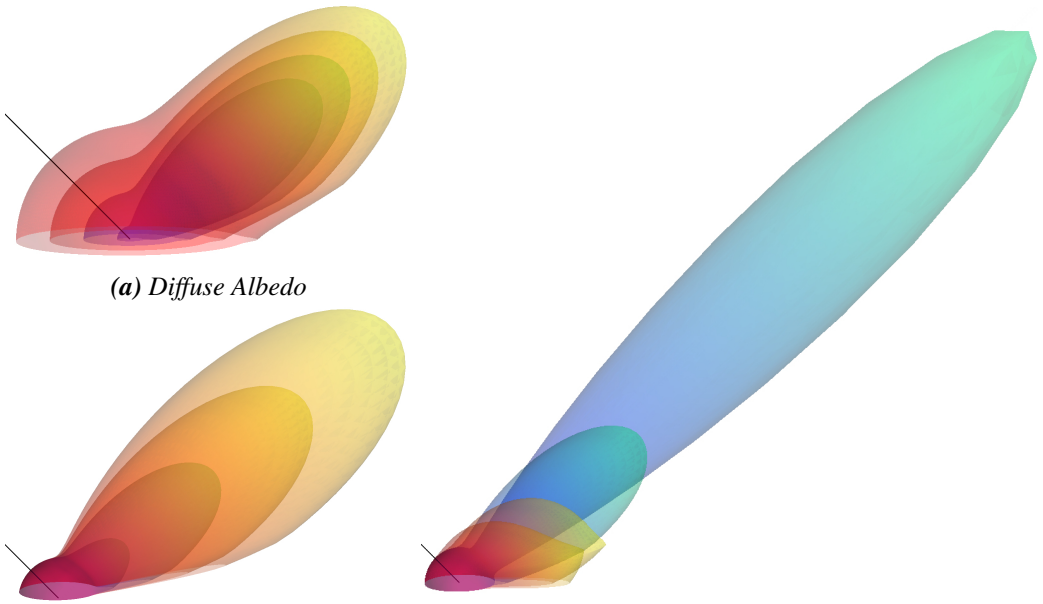
2.1.4 Parametric BRDF Models

There are two approaches to parametric modeling of BRDFs: Physically-based and phenomenological. Phenomenological models are designed to model observed scattering phenomena. As they are not necessarily created to follow physics, they do not always fulfill energy conservation, Eq. (2.19). Examples are the Phong model [Pho75], its extension by Blinn [Bli77], and the Lafortune model [Laf+97].

Physically-based models simulate physical interactions between light and materials. They usually strive for interpretable and manipulable parameters, and can lead to a good level of photorealism. Examples are the models by Torrance-Sparrow [TS67], Cook-Torrance [CT82], Oren-Nayar [ON94], and Disney [Bur12]. As most common parametric BRDF models, these are all so-called *microfacet models*. We will discuss the Cook-Torrance model as an example in the following.

Microfacet Theory: To explain reflective scattering phenomena, *microfacets theory* assumes that a surface consists of many microscopic mirror facets oriented following a probabilistic distribution function. As each microfacet reflects light in the mirror direction, a rough or diffuse surface is modeled by assigning a more complex geometry to the surface. Note that, in theory, microfacets are only appropriate when the facet areas are large compared to the wavelength of the incident light. However, it holds for most materials and works very well in practice. There are three terms to describe the microfacet model, which we introduce next.

1. Microfacet Distribution: The distribution of microfacets is described by the number of



(a) *Diffuse Albedo*

(b) *Specular Albedo*

(c) *Roughness*

Figure 2.6: Parametric BRDF Model. *Visualizations for different values of the (a) roughness, (b) diffuse albedo, and (c) specular albedo BRDF parameters (while fixing all other parameters) for the Cook-Torrance model. BRDF values are shown as in Fig. 2.5.*

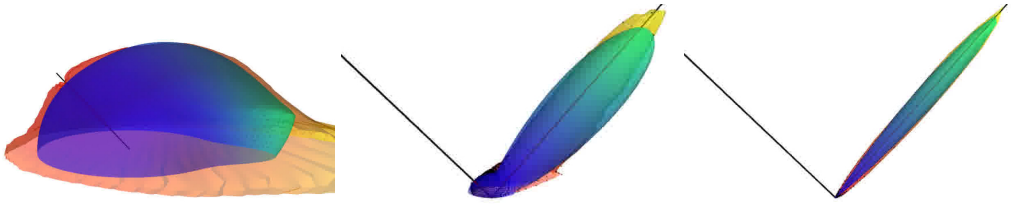
facets whose normals $\mathbf{n}_f$ form a certain angle α with the surface normal $\mathbf{n}$. Therefore, it is symmetric around the surface normal and implicitly assumes isotropy. Multiple formulations exist for this *microfacet slope distribution function* D. [Bur12] presents a generalization, the *Generalized-Trowbridge-Reitz (GTR)* distribution:

$$D_\gamma(\mathbf{n},\omega_{\text{in}},\omega_{\text{out}},r) = \frac{c}{\pi\left(\sin^2\theta_{\text{h}} + r^4\cos^2\theta_{\text{h}}\right)^\gamma}, \qquad c = \begin{cases} \frac{r^4-1}{\log(r^4)}, & \text{if } \gamma = 1 \\ \frac{(\gamma-1)(r^4-1)}{1-r^{4(1-\gamma)}}, & \text{else} \end{cases} \tag{2.21}$$

with roughness $r \in \mathbb{R}$, parameter $\gamma \in \mathbb{R}$, $\theta_{\text{h}} = \angle(\mathbf{n},\mathbf{h})$, and the half-way vector $\mathbf{h} = \frac{\omega_{\text{in}}+\omega_{\text{out}}}{|\omega_{\text{in}}+\omega_{\text{out}}|}$ between ω_{in} and ω_{out}. Visually, a distribution function with heavier tails (a larger γ) reveals more realistic results, [Bur12] suggest values of $\gamma \in [1,2]$.

2. Microfacet Shading: Modeling the surface as a collection of tiny microfacets causes masking and shadowing effects between the facets. Since all the facets are assumed to lie in a 2D plane, these only depend on the normals. A simple geometry factor accounts for that, the *geometric attenuation factor* G. The Mitsuba renderer [Jak10] uses:

$$G(\mathbf{n},\omega_{\text{in}},\omega_{\text{out}},r) \quad = \quad G_1(\angle(\mathbf{n},\omega_{\text{in}}),r^2) \cdot G_1(\angle(\mathbf{n},\omega_{\text{out}}),r^2) \tag{2.22}$$



(a) Red Fabric *(b) Cherry* *(c) Color Changing Paint 2*

Figure 2.7: Parametric Fitting to Measurements. *Fitting parameters of the Cook-Torrance model, Eq. (2.25), to measured BRDF data from the MERL database [Mat+03]. Visualizations show BRDF values as in Fig. 2.5; red/yellow colors show measurements and blue/green colors show the parametric fit. While not all details are recovered, approximations are reasonable.*

with a '*fast and accurate approximation*' to Smith's shadowing-masking function G_1:

$$G_1(\theta, r) \quad = \quad \frac{3.535a + 2.181a^2}{1 + 2.276a + 2.577a^2} \quad \text{and} \quad a = \frac{1}{r \tan\theta} \tag{2.23}$$

The Filament renderer [GA18] and [Ham17] extend Smith's GGX formulation and propose the height-correlated Smith visibility function:

$$G(\mathbf{n}, \omega_{\text{in}}, \omega_{\text{out}}, r) = \frac{0.5}{\cos(\theta_{\text{i}})\sqrt{r^2 + (1-r^2)\cos^2\theta_{\text{o}}} + \cos(\theta_{\text{o}})\sqrt{r^2 + (1-r^2)\cos^2\theta_{\text{i}}}} \tag{2.24}$$

for $\theta_{\text{i}} = \angle(\mathbf{n}, \omega_{\text{in}})$ and $\theta_{\text{o}} = \angle(\mathbf{n}, \omega_{\text{out}})$.

3. Fresnel Effect: Any electromagnetic wave, such as light, slows down and bends when entering optically denser material at an arbitrary angle. E.g., water becomes mirror-like when the light reaches a grazing angle as for a sunset over the sea. This is called the *Fresnel effect*. Here, the velocity, wavelength, propagation direction, as well as reflected and transmitted radiance of the incoming light changes (only the frequency stays constant). Generally, this effect occurs at all scattering events on surfaces, but as it depends on the index of refraction of both surface materials, it is mostly much less than it is for water or glass. Additionally, the effects are relevant near grazing angles only, and with an active handheld illumination setup, the Fresnel effect cannot be observed.

Cook-Torrance Model: The Cook-Torrance model [CT82] provides a good balance of expressiveness and compactness. It models reflectance as a sum of purely diffuse (i.e., Lambertian) and specular reflectance. For a point p with normal $\mathbf{n}_p$ the full formulation is given by

$$\rho(\mathbf{n}_p, \omega_{\text{in}}, \omega_{\text{out}}) = \frac{\mathbf{d}}{\pi} + s \cdot \frac{D(\mathbf{n}_p, \omega_{\text{in}}, \omega_{\text{out}}, r) G(\mathbf{n}_p, \omega_{\text{in}}, \omega_{\text{out}}, r) F(\mathbf{n}_p, \omega_{\text{in}}, \omega_{\text{out}})}{4(\mathbf{n}_p \cdot \omega_{\text{in}})(\mathbf{n}_p \cdot \omega_{\text{out}})} \tag{2.25}$$

with parameters: Diffuse albedo $\mathbf{d} \in \mathbb{R}^3$, specular albedo $s \in \mathbb{R}$, and surface roughness $r \in \mathbb{R}$.

Note: As done in prior work [Nam+18], we assume $F = 1$ in this dissertation, since it cannot be observed with our capture system. See a visualization of the parameter influences in Fig. 2.6 and fits of the Cook-Torrance model to MERL database BRDF measurements are shown in Fig. 2.7.

Spectral Distribution of Reflection: Up until this point, we have covered how to estimate the directional distribution of reflection. This means we know how to model that a couch looks similar from all angles and that a mug shows glossy highlights. However, we still need to determine how to model the spectral distribution and colors. A couch with a blue appearance means that light in the blue wavelength is reflected, but other wavelengths are absorbed. A mirror does not appear in any color since it reflects all wavelengths equally. Colors can be represented by many different models; a very common approach in Vision and Graphics is the RGB spectrum. It is straightforward and models the BRDF separately per color channel for red, green, and blue. We use the RGB color spectrum in this dissertation.

2.1.5 Forward Rendering

At the heart of modern rendering engines is the *rendering equation* or *light transport equation*, introduced by Kajiya in 1986 [Kaj86]. It describes the physical light transport path at each surface point p with surface normal $\mathbf{n}_p$ by relating incident radiance $L_{\text{in}}(p, \omega_{\text{in}})$, general surface reflectance $\hat{\rho}(p, \omega_{\text{in}}, \omega_{\text{out}})$, emitted radiance $L_{\text{e}}(p, \omega_{\text{out}})$, and outgoing radiance $L_{\text{out}}(p, \omega_{\text{out}})$ in an integral equation:

$$L_{\text{out}}(p, \omega_{\text{out}}) = L_{\text{e}}(p, \omega_{\text{out}}) + \int_{\mathcal{S}^2} \hat{\rho}(p, \omega_{\text{in}}, \omega_{\text{out}}) L_{\text{in}}(p, \omega_{\text{in}}) |\mathbf{n}_p \cdot \omega_{\text{in}}| \, \mathrm{d}\omega_{\text{in}} \tag{2.26}$$

This is a generalization of Eq. (2.20) to describe the full light scattering at a surface, including emitting surfaces and possible transmission by integration over the full unit sphere. The rendering equation Eq. (2.26) has no analytical solution. Due to object boundaries and visibility constraints, the integrand is discontinuous and potentially high-dimensional. Yet, there are numerical estimators, one of which we present next.

Monte-Carlo (MC) Estimator

Monte-Carlo (MC) integration methods provide one solution to approximating integral light transport equations $\int f(p)\,\mathrm{d}p$. MC integration methods use random sampling, evaluate the integrand $f(p)$ at the sample positions and compute an estimate for $\int f(p)\,\mathrm{d}p$.

Assume we want to approximate the integral $I = \int_a^b f(p)\,\mathrm{d}p$ and we are given a supply of random variables $X_i \in [a, b]$ with *Probability Density Function (PDF)* $q(p)$ that fulfills $q(p) \neq 0 \; \forall p : |f(p)| > 0$. Then the Monte-Carlo estimator

$$F_N = \frac{1}{N} \sum_{i=1}^{N} \frac{f(X_i)}{q(X_i)} \tag{2.27}$$

equals the integral I in expectation: $E[F_N] = I$, see [Vea98]. The estimator can be extended to higher dimensions by independently sampling each dimension and applying Eq. (2.27).

Therefore, applying MC integration to the rendering equation is straightforward. We assume the surface does not emit any additional light, i.e., $L_e(p, \omega_{out}) = 0 \, \forall p \, \forall \omega_{out}$. Then the Monte-Carlo estimator for Eq. (2.26) yields

$$L_{out}(p, \omega_{out}) \approx \frac{1}{N} \sum_{j=1}^{N} \frac{\rho(p, \omega_j, \omega_{out}) L_{in}(p, \omega_j) |\mathbf{n}_p \cdot \omega_j|}{q(\omega_j)} \tag{2.28}$$

for directions $\omega_j \in \{1, \ldots, N\}$ sampled according to some probability distribution over the unit hemisphere $\mathcal{S}^2_+$ with PDF $q(\omega_j)$.

Properties of MC Integration: For N samples, MC integration converges at a rate of $\mathcal{O}(N^{-1/2})$, independent of the dimensionality and smoothness of the integrand. Please see, e.g., [Vea98] for the derivation. A convergence rate independent of the integrand is the big advantage of MC integration. Computing the light transport for a scene means integration over all possible light paths - which are, due to depth discontinuities and visibility constraints, very discontinuous and, technically, this domain is infinite-dimensional. MC provides a solution for approximation in the form of an easy algorithm that solely requires sampling and the ability to evaluate the integrand as premises. The drawback of MC methods is the number of samples needed. Following the convergence rate, it requires $4N$ samples to cut the error in half. This is computationally expensive when rendering images with millions of pixels. Too few MC samples are typically visible as pixel predictions that are too bright or dark, showing in renderings as noise. Therefore, we briefly touch upon sampling next.

Generating Samples: Review Eq. (2.28): A remaining question is what probability distribution to sample from. Ideally, we would like to draw samples from multiple probability distributions, i.e., distributions that approximate the BRDF, the light sources, the camera, and possibly scattering media. There are multiple different sampling methods and variance reduction techniques. Especially for efficiency and rendering time, sampling is of great importance. Unfortunately, a detailed overview is out of scope for this work, and we refer the reader to [PJH16] for a thorough discussion on sampling techniques for rendering.

Rendering Techniques

There are multiple rendering techniques available, including rasterization, path tracing, ray tracing, sphere tracing, and bidirectional methods. Additionally, methods are classified as surface, volume rendering, or hybrid approaches. Due to the vast amount of information available, we cannot cover them in this dissertation. For a comprehensive introduction to existing rendering techniques, we recommend referring to sources such as [PJH16; Hug+13].

2.2 Scene Representations

A real-world scene is a complex arrangement of objects, materials, physical forces, light sources, optical phenomena, and much more. Modeling it from captures of optical sensing systems is challenging. One key question is how to represent this scene virtually in a machine-friendly way. Scene representations can be categorized into 2D (Sec. 2.2.1) and 3D, with the latter further divided into surface (Sec. 2.2.2) and volume (Sec. 2.2.3) representations. Surface representations encode properties of the surface; they can not capture volumetric matter like smoke. Volume representations, in turn, encode surfaces only implicitly as they represent information about the entire 3D space. Scene representations can also be characterized by whether they are discrete or continuous and whether they contain connectivity information. To provide a better understanding of scene representations, we present an overview of common scene representations in the following. For a more extensive discussion of scene representations in the context of neural rendering, readers can refer to the report in [Tew+22].

2.2.1 2D Representations

Typically, 2D scene representations are closely linked to the capture system, which returns measurements per pixel/unit of the electrical sensor. Examples include depth maps, disparity maps, and height maps. Note that depth maps are also often referred to as 2.5D since they allow for projection to 3D scene points and thus provide partial, view-centric 3D information. An advantage of representations in 2D is that images are trivially encoded in arrays or tensors, which allows for very efficient and fast calculations for millions of pixels simultaneously. Additionally, while they lack 3D connectivity, 2D connectivity information is encoded via neighboring pixels. A disadvantage of 2D representations is that their discrete nature constrains them to a fixed resolution where free space takes as much memory as any scene point. Further, since each pixel depicts a differently sized patch of the scene based on the orientation and distance of the camera, the representation does not allow to adjust memory to scene content.

2.2.2 3D Surface Representations

Surface representations can be explicit or implicit, referring to whether or not the surface is directly indexable. As explicit representations allow for direct extraction of surface points, they enable forward rendering into cameras via, i.e., rasterization. In contrast, rendering implicit representations requires sampling the 3D space, e.g., by ray casting.

Explicit Surface Representations: Explicit 3D representations include point clouds, meshes, and explicit surface functions. **Point clouds** naturally evolve from feature matching approaches. They are trivially spatially adaptive to the local resolution of the surface geometry. However, additional post-processing steps are required to extract a connected 3D surface. Also, no connectivity structure between points is contained, and the number of points needed to encode large scenes densely results in large memory footprints. **Meshes** are probably the most intuitive and best human-interpretable visual representation of 3D

surfaces. They provide connectivity information and enable real-time rasterization. In the context of 3D reconstruction, meshes tend to cause self-intersection and non-watertight surfaces. This can be prevented by transforming a template mesh, which, in turn, limits topological changes to sequential re-meshing. While points or mesh vertices are discrete, **surface functions** encode a continuous representation of the surface geometry. A common approach is the definition of a mixture of polynomial functions. While these work well for primitives, they are rather impractical for reconstructions tasks of scenes with unknown geometry. Another approach is to represent shape as a vector field and learn the thus defined diffeomorphism between a canonical surface, e.g., a sphere, and the target shape by a neural network.

Implicit Surface Representations: Implicit functions represent a surface as the level set or isosurface of a 3D function f: $\{x \in \mathbb{R}^3 | \, f(x) = c\}$ for $c \in \mathbb{R}$. Examples are Signed Distance Functions or occupancy functions. **Signed Distance Functions (SDFs)** store the shortest distance of each point to the surface, encode the inside and outside wrt. the surface by the sign and have the nice property that the gradient of the SDF aligns with the surface normal. **Occupancy functions** encode the inside and outside of a surface by assigning a binary value or occupancy probability to each point; the surface is then represented as the decision boundary between empty and occupied. All implicit surfaces are represented via discrete samples, stored in, e.g., a voxel grid, or encoded as a continuous function in a neural network. This way, they can encode arbitrary topologies. Extracting the surface from an implicit function can be expensive as it typically involves ray tracing or sphere tracing.

2.2.3 3D Volume Representations

Volume Representations encode the whole 3D space and are thus well suited to model participating media (like smoke or dust) or non-opaque objects (e.g., transparency). They can encode volume density, color, and any kind of features and are most often encoded in voxel grids or neural networks. Generally, images can be rendered by composition along rays. This is differentiable and works well in gradient-based optimization. However, volumetric representations usually depend on a heuristic threshold for surface extraction.

Voxels: Voxel grids are the direct generalization of pixels to 3D. Like pixels, they easily allow for computations as matrices or tensors. Yet, while their simplicity is attractive for many applications, the discretization of the full 3D domain yields a high memory footprint that scales cubically with the resolution. To reduce memory requirements, data-adaptive implementations, e.g., Octrees [LK11], enable efficient space partitioning.

Volumetric Neural Representations: Theoretically, any scene representation could be encoded in a neural network. Especially since MLPs act as universal function approximators [HSW89], scene representations that are mathematical functions are preferred in practice. Examples are density fields representing volume density per 3D point, radiance, or any feature vector of interest. The key benefit of a neural representation is that it is agnostic to grid resolution; it encodes a continuous 3D function that can be queried at arbitrary resolution. However, sampling can be expensive as each sample requires a forward pass through the network.

3 Related Work

We now discuss the most related work on geometry estimation in Sec. 3.1, material estimation in Sec. 3.2 as well as joint geometry and material estimation in Sec. 3.3. Further, we provide an overview on differentiable rendering approaches in Sec. 3.4 and 3D reconstruction at scale in Sec. 3.5.

3.1 Geometry Estimation

Geometry estimation from 2D images is a long-studied problem in Computer Vision. Early approaches lever either multi-view images and geometric cues, presented in Sec. 3.1.1, or multi-light images and shading cues, introduced in Sec. 3.1.2 and Sec. 3.1.3. Most newer works represent geometry implicitly and reconstruct geometry alongside lighting from multi-view images, covered in Sec. 3.1.4, or combine the benefits of previous works and reconstruct geometry from both multi-light and multi-view input images, which we elaborate on in Sec. 3.1.5.

3.1.1 Multi-View Stereo (MVS)

Multi-View Stereo (MVS) reconstruction techniques [SD97; Laf+13; UGB15; VTC05; KZ02; FP10; Sch+16; Zha+23d; ZZL23] recover the 3D geometry of an object from multiple input images by matching feature correspondences across views or by optimizing photo-consistency. As they ignore physical light transport, they cannot recover material properties. Furthermore, they are only able to recover geometry for sufficiently textured surfaces.

3.1.2 Shape from Shading (SfS)

Shape from Shading (SfS) techniques exploit shading cues for reconstructing [IH81; Zha+99; Hor70; Qué+17a; Qué+17b] or for refining [Hae+18; Wu+14; Zol+15; Mai+17] 3D geometry from one or multiple images by relating surface normals to image intensities through Lambert's law, Eq. (2.10). While early SfS approaches were restricted to objects made of a single Lambertian material, later reincarnations of these models [BM15; ON14; LN12] are also able to infer non-Lambertian materials and lighting. Unfortunately, reconstructing geometry from a single image is a highly ill-posed problem, requiring strong assumptions about the surface geometry. Moreover, textured objects often cause ambiguities as changes in either surface orientation or surface albedo can cause intensity changes.

3.1.3 Photometric Stereo (PS)

Photometric Stereo (PS) approaches [Woo80; QMD16; QLD15b; PF14; Hol+08; ZT10; Tun+13] assume three or more images captured with a static camera while varying illumination or object pose [Lim+05; SFB03] to resolve the ambiguities mentioned above. In contrast to early PS approaches, which often assume orthographic cameras and distant light sources, newer works consider the more practical setup of near light sources [LMC17; QWC17; XDW15; LND18] and perspective projection [Mec+14b; Qué+18; Mec+14a]. To handle non-Lambertian surfaces, [Qué+17c; QLD15a] suggest robust error functions, [MRC15; MQ16; Mec+16; CBR11] formulate the problem using specularity-invariant image ratios, and [Li+23] explicitly model shadowing and anisotropic reflectance. However, many classical PS approaches are not capable of estimating material properties other than albedo and most PS approaches require a fixed camera which restricts their applicability to lab environments. In contrast, here we are interested in recovering the shape and surface materials of larger scenes using a *handheld* mobile scanner.

3.1.4 Reconstructing Geometry Implicitly

Another line of works represents geometry implicitly as occupancy or density volumes and estimates these jointly with the appearance or radiance of each point given multi-view observations. Effectively, each 3D point is modeled as a small light source, like an emitting surface or surface light field [Woo+00]. Early examples modeling implicit surfaces or volumes are [Nie+20; Yar+20; Wei+21; OPG21; Yar+21; Wan+21a]. By using differentiable volumetric rendering and a photometric loss, they show appealing results for geometry reconstruction. However, since these initial models rely on a fixed 3D volume and do not model the physical light transport, they are not scalable and do not allow for recovering material properties. Such implicit, and foremost neural, representations form a very active field of research, and new models are developed rapidly for many different tasks. Attempting to give an overview in this dissertation is out of scope. Instead, we present a selection of works that extend implicit neural scene representations to larger scenes or include reflectance models in Sec. 3.4.2 and Sec. 3.5.1.

3.1.5 Multi-View Photometric Stereo (MVPS)

The advantages of PS (accurate normals) and MVS (global geometry) have also been combined by integrating normals from PS and geometry from MVS [EVC08; Lu+10; Neh+05; JK07; Fan+18; Par+17; Shi+14; YY11; LMC19] into a single consistent reconstruction. Most recent approaches [ASH22; Kay+22; Yan+22; Kay+23] use volume rendering of implicit 3D representations to reconstruct geometry. As inference of such representations is expensive, [Zha+23e] predict depth maps from cost volumes of aggregated feature maps. Many of these latest neural network-based methods take the view direction into account and exhibit robustness to non-Lambertian specular reflectance. Yet, all require multi-light images from a fixed camera position as input and suffer from slightly blurry predictions around geometric details. In contrast, we target accurate geometry and material reconstructions from handheld data.

3.2 Material Estimation

The previously discussed geometry reconstruction methods in Sec. 3.1 model surface appearance with textures that store color or albedo. This allows for re-renderings of Lambertian surfaces but does not capture view- or illumination-dependent effects, thus, resulting in poor re-renderings when changing the viewpoint or illumination. In Sec. 3.2.1, we present Intrinsic Images, the earliest approaches for separating view-dependent and view-independent properties given a single image.

The Bidirectional Reflectance Distribution Function (BRDF), [Nic+92], offers a more precise representation; we have presented it in Sec. 2.1.2. For a single material, we have discussed data-driven and parametric models of the reflectance function in Sec. 2.1.3 and Sec. 2.1.4. Now, we discuss how to estimate the reflectance properties from images and compose a surface of multiple materials in Sec. 3.2.2. Note: With the term *material estimation*, we refer to estimating the reflectance function, a material property.

3.2.1 Intrinsic Image Decomposition

Intrinsic Image Decomposition [Bar78] is the problem of decomposing a single image $\mathcal{I} \in \mathbb{R}^{H \times W \times 3}$ into its material-dependent and light-dependent properties:

$$\mathcal{I} = \mathcal{R} \cdot \mathcal{S} \tag{3.1}$$

Hereby, material properties are modeled by a reflectance image $\mathcal{R} \in \mathbb{R}^{H \times W \times 3}$, depicting the light-independent intrinsic color of a surface, and a shading image $\mathcal{S} \in \mathbb{R}^{H \times W \times 1}$ that captures the light-dependent effects. This problem is fundamentally under-constraint; strong regularizers must be exploited to constrain the solution space and reduce ambiguity. A common assumption is to model all reflections as Lambertian.

The first influential solution to the problem was the Retinex method [LM71] from 1971, which assumes that shading causes small image gradients, whereas reflectance leads to bigger changes in image intensity. It has led the top-performing approaches for almost four decades, [Gro+09]. More recent approaches rely on global sparsity constraints and priors [Geh+11; SYH13; STL08; Kov+17; LS18a; Liu+20c], assume depth image as additional input [CK13; Xin+23], leverage explicit geometric constraints by additionally reconstructing geometry [BM15; Kim+16; Jan+17], or learn to model non-Lambertian specular highlights [Shi+17]. Further extensions of the Intrinsic Image problem exploit multi-light images [LS18b; Liu+20a] to learn reflectance and shading decomposition unsupervised or relax the assumption of a single input image and leverage multiple images to guide the decomposition [Laf+12; Mél+18].

However, while the decomposition into shading and reflectance images proved beneficial for related tasks such as geometry reconstruction, all the models only capture a small portion of the 3D physical process that forms an image. This simplistic image formation model does not model many effects that occur when light hits a surface. This leads to artifacts in the predictions and does not allow for realistic renderings under changed viewpoints or illumination. A more accurate description of reflectance properties is discussed next.

3.2.2 svBRDF Estimation

Estimating the reflectance properties of real-world materials poses two major challenges: First, surface appearance depends on the orientation of the surface. Especially, small geometric variations (texture in the '*non-graphics*' sense) like scratches or dents (e.g., objects warn off from use) vary the appearance greatly. Therefore, knowledge of the scene geometry is crucial for accurate material reconstruction. Second, real objects, let alone scenes, are composed of multiple materials, each of which exhibits lots of tiny variations. Dust, stains, or fat occur naturally from age or touch and locally change the BRDF. Therefore, most realistic reflectance representations are *spatially-varying*, meaning that they model per-point reflectance properties and, thus, can depict all local changes.

In the following, we focus on methods that recover *spatially-varying Bidirectional Reflectance Distribution Functions (svBRDFs)*. We categorize them by their priors on geometry as well as the practicality of the acquisition system.

Known Geometry: Knowing the 3D geometry (e.g., accurately measured or using primitives), the BRDF can be inferred using specialized light stages or gantries [Mat+03; Len+03; Law+06; Sch+13; ISS20; NJR15]. While this setup leads to accurate reflectance estimates, it is typically expensive and requires millions of samples. More lightweight approaches reconstruct BRDFs from sparse but accurate measurements of planar surfaces [Xu+16], or use consumer-grade capture setups for planar [AWL13; Ren+11; ZK21] or complex objects [Don+14; Zho+16; Gho+09]. All these provide dense samples but require stationary capture setups or static cameras and only work for objects of limited size.

To simplify the acquisition system drastically, multiple works have demonstrated that reflectance properties can be acquired using a handheld camera, e.g., a mobile phone, and multiple images [Hui+17; RPG16; AWL15; Gao+19; Alb+18], or even a single image [Des+18; LSC18; ZK22; Guo+23; Zha+23b; Wen+22; SLS23]. Since static illumination, or a single, co-located flashlight is used, this allows for very sparse samples only. For more dense BRDF samples, [ALL20] propose a portable multi-light capture apparatus. While data collection is easy and practical, these techniques are designed for capturing flat textured surfaces and do not generalize well to objects with more complex geometries.

Unknown Geometry: More closely aligned with our goals are approaches that estimate parametric BRDF models for objects or scenes with unknown geometry from sparse measurements: [Hae+21] use precomputed geometry from the Replica dataset [Str+19], [WZ15; WYT16; WWZ16; PNS18; PHS20] calculate geometry from RGB-D data of commodity RGB-D sensors using KinectFusion [New+11; Iza+11], [Yu+99; Kim+17; Yao+22; JP22; WJP21] infer geometry from RGB input applying MVS methods, and [Pra+22] use a precomputed MVS mesh that gets refined and cleaned by an artist. These methods infer geometry first and calculate reflectance in a subsequent step.

Recall the beginning of this section: Reflected appearance is very susceptible to small geometric variations and shading cues are beneficial to reconstruct geometry accurately. Therefore, we *jointly* optimize for poses, geometry, and material parameters. Our experiments in Sec. 4.6.3 confirm that joint optimization is superior over a corresponding disjoint formulation: It recovers fine geometric structures while refining material estimates simultaneously.

3.3 Joint Geometry and Material Estimation

Several works have addressed the problem of jointly inferring geometry and materials. By integrating shading cues with multi-view constraints and accurate models of materials and light transport, this approach has the potential to deliver the most accurate results. However, joint optimization of all relevant quantities is a challenging task. We distinguish between static (Sec. 3.3.1) and mobile (Sec. 3.3.2) acquisition systems in the following.

3.3.1 Static Capture Systems

Accurate 3D models of both shape and reflectance properties of single objects can be acquired with sophisticated scanning systems and light stages [Kan+19; HLZ10; Tun+13; Riv+20]. For more lightweight approaches, several works have considered extensions of the classic PS setting to predict shape and spatially-varying BRDFs [Gol+10; AZK08; HS05; HS15; HS17; Lic+21] and to multiple viewpoints and 3D models [Bir+06; ZWT13; Li+20a]. Relaxing the PS-constraint, [Xia+16] use alternating optimization, and [DLG21; Bi+20c] use multi-branch encoder-decoder architectures to estimate geometry, surface normals, and spatially-varying BRDFs. But they all require multiple images from the **same or known viewpoints** as input. In contrast, we are interested in jointly estimating geometry and materials from *mobile scanning systems*, enabling applications outside laboratory environments.

3.3.2 Mobile Scanning Systems

In 2011, [Ren+11] proposed to exploit low-cost and handheld scanning devices such as a flash camera for reconstructing BRDFs and geometry from multi-view images. Like some subsequent works [Kim+17; Che+21a; Hig+09; Mai+17], they are restricted to the Lambertian reflectance model or uniform materials.

For **single objects or small scenes** many recent approaches propose to train neural networks to reconstruct an implicit representation of geometry alongside svBRDF parameters. While [Li+22c; Bos+20] train their models on synthetic data for 2.5D predictions, [Bi+20a; Sri+21; Ver+22; Kua+22; Bos+21a; Bos+21b; Zha+21c; Bos+22; Bi+20b; Che+21b; Zha+21b; YN23; Mao+23; Fan+23; Zha+23c; Mai+23; Zha+23a; Wu+23a] train the proposed models unsupervised per object in 3D. These works rely on fully-connected network architectures to represent the scene. While MLPs are simple and powerful, these models are designed for single objects and do not scale beyond, see [Pen+20]. Further, inference is expensive, and reconstructions typically suffer from over-smoothing and a lower accuracy for geometric details.

To alleviate these last two aspects, the following ideas were proposed very recently: [Sun+23b; Sun+23a] use learned neural representations and volumetric rendering for continuous 3D feature extraction and then optimize for non-neural scene parameters using a physically-based surface renderer based on these features. Similarly, [Wan+23] combines volume rendering of a neural field with ray-tracing of an explicit mesh. [Zha+23c; Mai+23] introduce microflake or microfacet fields: Instead of modeling each point in a volume as an

emitting particle (as done in radiance fields, e.g., NeRF), they compose the volume of small oriented micro surfaces and show impressive results for volumetric matter (e.g., clouds) and fine geometric details. All these models are designed for objects only and do not easily scale beyond – which is a goal of the work presented in this dissertation. Nevertheless, they pose exciting opportunities for future development, which we come back to in Chap. 6.

Without relying on neural networks, the following works estimate materials alongside 3D geometry: Georgoulis et al. [GPG14] optimize 3D geometry and a data-driven BRDF model alternatingly. Nam et al. [Nam+18] refine a subdivided mesh by alternatively updating positions, normals, and material properties. Last, Li et al. [Li+21a] iteratively optimize for 3D geometry, reflectance, camera pose, and environment lighting. All these methods decompose the problem into smaller problems by splitting the optimization variables by their property (i.e., geometry, materials, poses) and alternate the optimization over those properties using multiple decoupled objectives.

In contrast, we exploit that spatially separated regions naturally decouple the corresponding optimization variables, and therefore, we decompose the problem based on spatial regions instead of separate properties. This has two advantages: 1) It enables us to optimize each region's parameters jointly and use a single objective function. Consequently, we can use all information encapsulated in the intricate interplay of geometry and materials and reach highly accurate reconstructions. 2) The separation into spatial regions allows us to distribute the optimization, which facilitates scalability to larger scenes.

Follow-up Work: Following our publication [Sch+20], Luan et al. [Lua+21] proposed another method for jointly optimizing geometry and spatially-varying reflectance. They represent the geometry of a single object as a mesh and alternate the optimization over mesh vertices and reflectance with re-meshing in a coarse-to-fine process. They use a co-located configuration of a handheld camera and point light which significantly simplifies the rendering process, yet, it restricts the sample space of the BRDF. In contrast, we use multiple light sources in conjunction with explicit shadow modeling to extract the most information from the sparse samples captured with a handheld setup.

With our proposed method, we make a step towards reconstructing geometry, materials, and poses beyond just the object-level. In contrast to methods that assume watertight 3D shapes, our model is able to reconstruct partially scanned 3D environments as it does not require closed shapes.

3.4 Differentiable Rendering

Differentiable rendering describes a rendering pipeline that allows for computing image pixel changes wrt. scene parameters. It is at the heart of most methods that aim to synthesize photo-realistic images from real-world observations. As [VSJ22] put it: '*Methods for physically-based differentiable rendering (PBDR) are of increasing interest due to their ability to solve previously intractable inverse problems involving realistic material appearance, shadowing and interreflection.*' The approaches are manifold, but many share a similar structure: Based on real 2D or 2.5D observations, an inverse renderer is used to infer a parametric and potentially neural representation of the 3D scene (e.g., for geometry, illumination, or BRDF). This scene representation is then rendered into images by a forward rendering engine. In this step, differentiable rendering enables the propagation of gradients to the parameters based on the rendered images. Differentiable renderers have been developed for various geometry representations, e.g., meshes [NJJ21], pointclouds [Yif+19; Wil+20], implicit surfaces [Nie+20], SDFs [VSJ22; Jia+20b], as well as rendering algorithms, e.g., sphere tracing [Liu+20b], or rasterization [KUH18; Liu+19].

Thereby, different approaches have emerged that we present in the following: *Inverse Rendering* methods (Sec. 3.4.1) use parametric scene representations plus classical forward renderers and optimize for, or learn, inferring the scene parameters from observations. *Neural Rendering* (Sec. 3.4.2) refers to methods that learn either the scene representation or the rendering engine by a neural network. Note that these distinctions are not exclusive: Most neural scene representation methods perform neural inverse rendering.

3.4.1 Inverse Rendering

Classical inverse rendering approaches use non-learning-based methods to optimize 2D or 3D scene parameters from observation images via gradient descent. [Liu+19; Rav+20] use *soft* rasterization to differentiate a rasterizer, while [LB14; LHJ19; Li+18; Nie+20; Lua+21; Nim+19; VSJ22] propose solutions to differentiate through ray casting. All these works use hand-designed rendering functions. This limits the flexibility of the renderer (as compared to learning-based approaches, discussed afterward) but has the advantage of physically-based rendering functions which are interpretable and enable rendering a scene under changed conditions (e.g., novel viewpoint, different illumination, or edited materials). In the pipelines presented in this dissertation, we use such a classical optimization approach with a hand-designed rendering engine since we aim for physically correct reconstructions.

Neural Inverse Rendering: Recent pipelines train neural networks to predict scene geometry [Kim+17; Wan+21b; LYC20; WWR22], and spatially-varying materials [LSC18; Des+18; Bos+20; Li+22d; Li+22c; Zhu+22b; Gao+19; ZK21; Lic+21; ALL20; Bi+20c; DLG21; Li+20b; Kan+19; ZK22; Guo+23; Zha+23b; Pra+22; SLS23; Wen+22; Zhu+22a] from observations. Then, as in classical inverse rendering, they use an analytical differentiable rendering layer to synthesize images. Here, the neural network has the potential to learn to ignore transient objects, adapt to varying illumination conditions in the observations, or disentangle the parameters of complex scenes from data. Yet, it also introduces additional parameters and, thus, requires data to train. This is not easy to obtain, especially when

considering more complex reflectance settings. In contrast, our approaches do not require any large dataset but solely takes the captures of a scene as input.

3.4.2 Neural Rendering

We now discuss neural scene representations and neural rendering approaches. Hereby, we follow the definitions in [Tew+22]: 2D neural rendering refers to methods that, given a parametric representation of the scene, use a generative neural network to model the rendering engine. 3D neural rendering or neural scene representations train a neural network to represent the 3D scene and combine these with a classical rendering engine.

2D Neural Rendering: Also called *neural rendering*. It refers to methods that use classical surface or volume representations and replace the differentiable rendering engine with a generative model to learn the image formation function. Exemplary tasks are changing the camera viewpoint [Esl+18; Mes+19; Thi+20; Ngu+18; Hed+18; Xu+19; Sit+19] or relighting [Gao+20; Xu+18; Phi+19]. The generative network has the potential to synthesize high-quality novel images, learn visibility constraints, deal with incomplete or inconsistent input representations, or depict complex illumination effects like inter-reflections or multiple bounces. However, the rendering network is non-deterministic, and how to enforce physical plausibility of the reconstructions is unclear – which is the goal of this dissertation. Therefore, the proposed approach relies on a classical rendering engine instead of a neural renderer.

3D Neural Rendering or Neural Scene Representations: These are methods that encode the scene in a neural network and combine this representation with classical differentiable rendering engines. A very well-known example are *Neural Radiance Fields (NeRF)* [Mil+20], which learn to encode continuous 3D fields for opacity and color in an MLP with a fixed scanning volume for novel view synthesis. Its follow-ups, e.g., [Zha+21b; Sha+21; Bos+21a; Zhu+23b] enable relighting and [Zha+21c; Sri+21; Bi+20a; Bos+21b; Bos+22; Kua+22; Mun+22; Zha+22; Yao+22; Zha+21a; Mao+23; Fan+23; Wu+23b; Zha+23c; Zha+23a; Mai+23; Wu+23a; YN23; Liu+23] include full BRDFs to encode material reflectance. These approaches are trained or fine-tuned unsupervised per scene, which, on its own, provides insufficient guidance for reliable reflectance estimation. Therefore, most models mentioned above either include prior knowledge of materials (e.g., via pre-trained reflectance or transmittance priors, using objects with very little material variation, or given a material segmentation) or they employ strong regularizers (e.g., compression to low-dimensional or sparse latent spaces, regularization against diffuse materials or a single specular material, or encoding in Spherical Gaussians). Exceptions are: [Zhu+23a] rely solely on regularization of smoothness and energy conservation, and [Had+23] introduce a new *neural radiometric prior* which regularizes the radiance field values directly against the rendering equation. Our method is also optimized per scene, yet, we do not assume any prior knowledge of materials. Additionally, our representation is faster to optimize than an MLP, naturally scales to large scenes, and distributes modeling capacity equally across the scene.

Instead of using radiance fields, [Bi+20b] uses a voxel representation of deep features that encodes opacity, normals, and materials. For a scene with multiple objects, this requires

compromising reconstruction resolution due to a fixed encoding which limits scalability. Neither using radiance fields nor storing deep features discretely, [Che+21b] track the 2D topology of the 3D surfaces of an object by optimizing the neural transform from the unit sphere to the 3D object. By design, this method re-trains the network individually for each object in an object-centered unit volume. This maximizes reconstruction quality per object but at the cost of a global scene representation.

In contrast to both methods, we propose a method the resolution does not depend on the scene size, and we optimize all local representations in a single global world coordinate system which does not require any reconfiguration of the representation after initialization.

3.5 3D Reconstruction at Scale

To date, there are few solutions to reconstructing accurate geometry and materials beyond object-scale. We review related work on scaling geometry estimation in Sec. 3.5.1, on room-scale material estimation in Sec. 3.5.2, and discuss existing approaches on geometry and material estimation of multi-object scenes in Sec. 3.5.3.

3.5.1 Scaling Geometry Estimation

In this paragraph, we review existing work on scalable geometry-only reconstruction. Crucial for a scalable model is a memory-efficient scene representation that allows for accurate and dense reconstructions. There are two alternatives for handling processing memory: Either the model must be very memory-efficient such that it fits in memory as a whole. Or, the model is designed to allow for division into parts or segments with a much-reduced memory footprint. We elaborate in the following.

Keeping the full Model in Memory: One approach is to keep the full reconstruction in memory by supporting efficient compression of connected surface data. [Hua+17; Laf+13] represent the scanned environment by a combination of meshes and geometric primitives, and [GPF10] fits a multi-layer height map to a volume of occupancy votes. These representations support scene completion and can scale efficiently to larger scenes but fail to reconstruct complex 3D structures with fine details. In the context of *Image Based Rendering (IBR)*, [Hed+16; Pra+21] calculate a global mesh but then refine per-view depth maps, sacrificing global consistency for local accuracy. [Bar+21; Bar+22; Zha+20] use neural radiance fields encoded in an MLP, and propose extensions of NeRF [Mil+20] to unbounded and larger-scale scenes. Similarly, [Azi+22; Dai+21] encode geometry in a *Truncated Signed Distance Function (TSDF)* represented by an MLP or volumetric grid. These volumetric representations scale with scene size and do not allow for modeling very large scenes. In contrast, the proposed method optimizes for a globally consistent and accurate scene representation that does not scale with scene size.

Models that allow for Subdivision: The alternative is to use scene representations that allow for subdivision into parts or segments and to introduce hierarchies to facilitate memory-efficient processing by keeping only relevant scene parts in memory. A common strategy is to first reconstruct the geometry of individual and overlapping scene segments, parts, or frames and then integrate these segments in 3D world space while employing sophisticated pose registration, alignment, and outlier filtering techniques. The main challenge is to achieve consistent and accurate global models based on local segments given data from handheld sensors.

Multiple *implicit volumetric models* extent the seminal works of [CL96; Lev+00] or KinectFusion [New+11; Iza+11] (which rely on memory-inefficient regular voxel grids) to larger environments by introducing efficient volumetric data structures like volume windows [Whe+12; RV12], patch volumes [Hen+13], a hierarchical volume structure [Zen+13; CBI13], or spatial hashing [Nie+13; Dai+17]. These increase spatial efficiency. [ZK13] increase spatial consistency and accuracy of large-scale reconstructions by preserving de-

tailed geometry around points of interest. Next to these SDF-based approaches, other works represent geometry as local implicit functions, encoded in voxel grids [Jia+20a], radiance fields [Tan+22; Zha+22; GCS23], or convolutional occupancy crops [Pen+20].

Non-volumetric approaches represent scene segments by, e.g., per frame 2.5D depth maps [Pol+08; DG19; Yan+20], 3D mesh fragments [CZK15; ZMK13; Whe+16], partial point clouds [Aga+09], or surfels [Bad+20].

All subdivision methods mentioned above enable final reconstructions that would exceed memory limitations during processing. And most employ global registration, pose refinements, or texture optimization like [ZK14; BKR17; Hua+17; Hua+20; CDK20; YYH20; Zen+17]. The common challenge of all non-global geometry reconstruction methods is ensuring consistency between local reconstructions. The method presented in this dissertation represent the scene as a collection of 2.5D parameter maps from multiple keyframe views. This representation is memory-efficient, and we actively encourage consistency between overlapping regions. As demonstrated in Sec. 4.6 and Sec. 5.6, this leads to accurate and well-aligned reconstructions, eliminating the need for post-processing or refinement.

Note: In this section, we focused on the memory efficiency of different methods since this is key to enabling scaling. We discuss additional challenges arising when scaling, e.g., temporal occluders or online image collections, in Sec. 6.4.1.

3.5.2 Room-Scale Material Estimation

Given Accurate Geometry and Poses: Assuming known geometry, camera poses, and object segmentation, [Yu+99; Azi+19; Hae+21; Nim+21] reconstruct reflectance properties of a full room. The earliest work from 1999 [Yu+99] assumes direct illumination from multiple light sources and shows results on real captures of a single room. The following, more recent works extend to arbitrary lighting. While [Azi+19] mainly targets synthetic data, [Hae+21; Nim+21] use geometry and segmentation provided by the *Replica* dataset [Str+19]. [Nim+21] explicitly models multiple bounces of light paths for rendering, but all these approaches model specular reflectance per object and not spatially-varying. These works leverage given segmentation and accurate geometry reconstructions and are not designed for imprecise inputs, like geometry reconstructions from handheld capture systems. Additionally, all but [Yu+99] (who estimate a BRDF for each surface independently) use global scene representations that are kept in memory during optimization. This limits scalability for the reasons discussed in Sec. 3.5.1.

Estimated Geometry and Poses: A few works tackle material estimation given geometry and poses estimated by classical, off-the-self MVS techniques. [Laf+12; LS18b] scale the intrinsic images problem to photo-collections of outdoor scenes and estimate shading and reflectance images under the Lambertian world assumption. Full svBRDFs are recovered by [JP22] for outdoor monuments with temporary occluders and by [Li+22b; Wu+23b] for indoor rooms. Here, [Li+22b] models multiple bounces by introducing a novel lighting representation based on *High Dynamic Range (HDR)* textures. [Wu+23b] use a factorized light transport formulation and detect emitters via rendering errors. As these works fix input

geometry and poses, they do not lever the information encapsulated in the interplay of reflectance estimates, geometric details, and poses.

3.5.3 Large Scale Geometry and Material Estimation

There exist only a few approaches tackling geometry and material reconstruction for scenes with multiple objects: [Li+21b; Li+20b; Zhu+22a; Li+22c; Phi+21; Zhu+23a; Wan+23; Wu+22].

Li et al. [Li+21b] tackle the challenging task of building a synthetic dataset of complex indoor scenes from commodity 3D sensor data. The dataset *OpenRooms* features geometry, svBRDF, and lighting of room-scale scenes rendered with up to seven bounces of light interactions per ray. While based on captures of real scenes, Li et al. do not do not pursue mimicking input appearance. They fit primitives and existing CAD models to the captured geometry, assign random materials per object, and rely on manual labeling and annotations to create a photo-realistic synthetic dataset.

In contrast, Li et al. [Li+20b], Zhu et al. [Zhu+22a], and Li et al. [Li+22c] predict 2.5D parameter maps from a single input image or panorama of an indoor room. All leverage multi-branch encoder-decoder networks with analytical physics-based rendering layers and recover depth and normal maps, spatially-varying BRDF parameters, as well as lighting. Yet, since all predictions are with respect to the input image, the results are limited by the single perspective and do not provide full 3D models.

Designed for novel view synthesis and relighting of full 3D rooms, Philip et al. [Phi+21] use MVS to reconstruct the scene geometry and then predict a diffuse albedo mesh, as well as view-dependent mirror images and view-independent irradiance images from RGB input. Using a mixture of image-based and physically-based rendering, they present convincing results for novel views of indoor rooms with multiple glossy surfaces and well-recovered specular highlights despite not modeling full BRDFs.

Zhu et al. [Zhu+23a] and Wang et al. [Wan+23] learn neural implicit fields encoded in separate MLPs to recover shape, material, and lighting from multi-view images. While the models are trained fully unsupervised, the fully connected networks struggle to recover high-frequency textures and fine details.

Very closely aligned with our goals is last year's work by Wu et al. [Wu+22]: They reconstruct a room's geometry and surface appearance under arbitrary, static illumination from RGB input images. Using MLPs to represent the scene, their volumetric approach can predict convincing relighting results. Similar to our method, Wu et al. propose a non-global scene representation and partition 3D space into tiles, each of which is assigned a surface and a reflection MLP. This representation facilitates scalability by avoiding training a large-capacity MLP, but the authors point out the need to ensure global consistency of reflection and geometry predictions across local tiles. We report a similar observation in Sec. 5 for our non-global keyframe-based representation. Unlike us, their method partitions the 3D volume, whereas we partition the 2D observations. While their approach avoids the need for sampling keyframes that we face, they require background sampling to calculate the photometric loss in image space. Additionally, they do not restrict lighting as we do. Instead,

they model the reflectance MLP with Spherical Harmonics, which introduces smoothing, and they separate reflections by modeling virtual light sources underneath the surface. While the latter is accurate for flat surfaces, it likely fails for non-planar glossy surfaces. In contrast, our approach restricts scene illumination but allows for more accurate geometry and material disentanglement for all surface shapes.

In conclusion, the method presented in this dissertation recovers real-world appearance by modeling 3D geometry and svBRDFs like none of the above does. We include a discussion on scalability in Sec. 6.4.

4 On Joint Estimation of Pose, Geometry and svBRDF from a Handheld Scanner

Towards 3D reconstructions that enable photo-realistic re-renderings, we first present a formulation for joint recovery of camera pose, object geometry, and spatially-varying BRDF in 2.5D. The input is a sequence of RGB-D images captured by a mobile, handheld scanner that actively illuminates the scene with point light sources. As such data is sparse, especially concerning material samples, the proposed method employs effective regularization to propagate material information across pixels and enable accurate recovery.

Further, the complex interplay of geometric and material entities makes the recovery and disentanglement of parameters very challenging. While previous works on geometry and material estimation from a handheld scanner tackle the problem sequentially or alternately, we formulate this problem using a single objective function that can be minimized using off-the-shelf gradient-based solvers.

This work is published in CVPR 2020, [Sch+20].

We start this chapter by describing the problem of joint geometry, material, and pose estimation from sparse input data in greater detail in Sec. 4.1. We then introduce our reconstruction algorithm: The scene representation is detailed in Sec. 4.2, the optimization objective is presented in Sec. 4.3, and we elaborate on our optimization in Sec. 4.4. Next, we show our custom-built handheld sensor and introduce the data acquisition pipeline in Sec. 4.5. Last, we present experimental evaluations of the proposed method in Sec. 4.6: We provide comparisons to various baselines and a study on the importance of each component in our formulation. E.g., we show that optimizing over the poses is crucial for accurately recovering fine details and that our approach naturally results in semantically meaningful material segmentation. Additionally, we demonstrate that by integrating material clustering as a differentiable operation into the optimization process, our model is able to recover the reflectance of specular materials independently. Results are presented for a diverse set of objects featuring various material variations and a small scene composed of multiple objects with strong occlusions and shadowing.

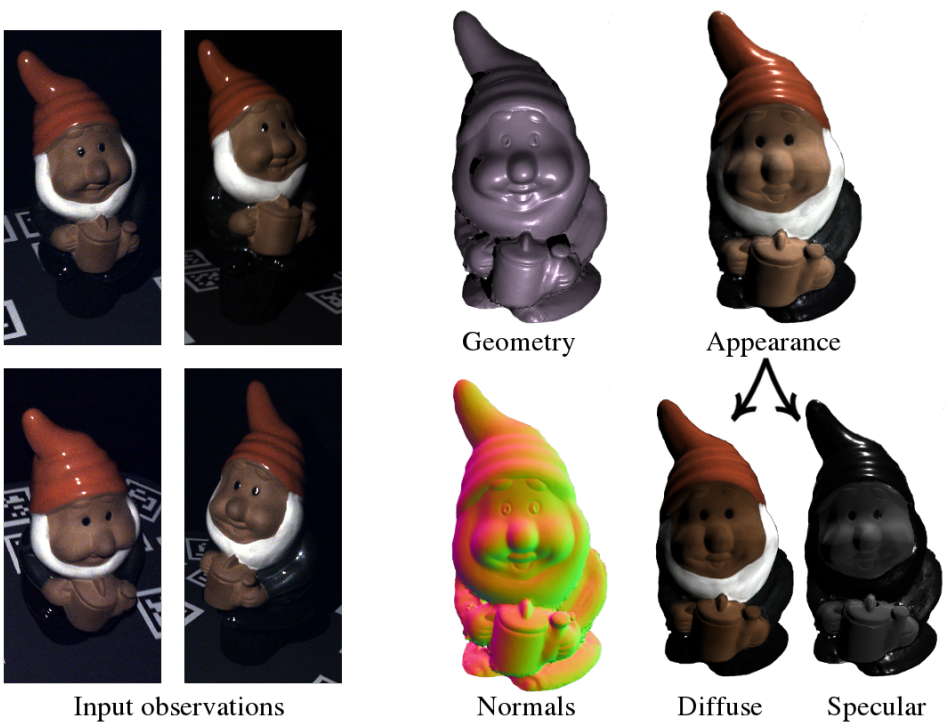


Figure 4.1: Method Overview (2.5D). *Based on images captured from a handheld scanner with point light illumination, we jointly optimize for the camera poses, the surface geometry, and spatially-varying materials using a single objective function.*

4.1 Introduction

Reconstructing the shape and appearance of objects is a long-standing goal in computer vision and graphics, with numerous applications ranging from telepresence to training embodied agents in photo-realistic environments. While depth sensing technology (e.g., Kinect) enabled large-scale 3D reconstructions [Nie+13; CZK15; Whe+15], the level of realism provided is limited since physical light transport is not taken into account. Consequently, material properties are not recovered, and illumination effects such as specular reflections or shadows are merged into the texture component.

Material properties can be directly measured using dedicated light stages [Mat+03; Len+03; HLZ10] or inferred from images by assuming known [Don+14; PNS18; Kim+17] or flat [Alb+18; Hui+17; RPG16; AWL15] object geometry. However, most setups are either restricted to lab environments, planar geometries, or difficult to employ *in the wild* as they assume aligned 3D models or scans.

Ideally, object geometry and material properties are inferred jointly: A good model of light transport allows for recovering geometric detail using shading cues. An accurate shape model, in turn, facilitates the estimation of material properties. This is particularly relevant for shiny surfaces where small changes in the geometry significantly impact the appearance

and location of specular reflections. Yet, joint optimization of these quantities (shown in Fig. 4.1) is challenging.

Existing approaches in 2020 address this problem by assuming multiple images from a static camera [Gol+10; HS05; HS17; EVC08; Zuo+17], which is impractical for mobile scanning applications. Only a few works in this time consider the challenging problem of joint geometry and material estimation from a handheld device [Hig+09; GPG14; Nam+18]. However, they assume known camera poses and leverage sophisticated pipelines, decomposing the problem into smaller problems using multiple decoupled objectives and optimization algorithms that treat geometry and materials separately. Furthermore, the number of base materials must be provided, and/or pre-processing is required to cluster the object surface accordingly.

At the time of publishing, we were the first to provide a formulation for this problem that does not rely on sophisticated pipelines or decoupled objective functions. Our method assumes the data was captured under known, non-static illumination with negligible ambient light. Our contributions are the following:

- We demonstrate that joint optimization of camera pose, object geometry, and materials is possible using a single objective function and off-the-shelf gradient-based solvers.
- We integrate material clustering as a differentiable operation into the optimization process by formulating non-local smoothness constraints.
- Our approach automatically determines the number of specular base materials during optimization, leading to parsimonious and semantically meaningful material assignments.
- We provide a study on the importance of each component in our formulation and a comparison to various baselines.
- Our source code, dataset, and reconstructed models are publicly available at https://github.com/autonomousvision/handheld_svbrdf_geometry/.

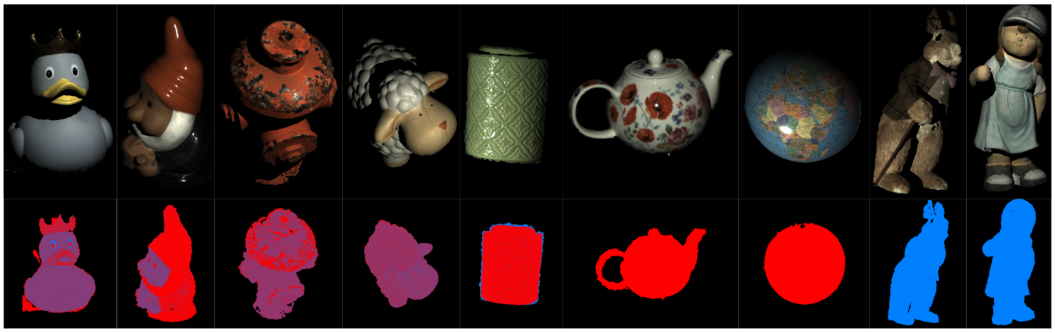


Figure 4.2: Reconstructions and Estimated Material Assignments for Real Objects (2.5D). *On test views, our rendered results (top) and specular base material segmentations (bottom) show accurate reconstructions for a variety of materials: The 'Teapot' and the 'Globe' are very smooth and reflective objects, whereas the 'Girl' and 'Rabbit' are near-Lambertian objects with rich geometry and a rough surface. The 'Duck', the 'Gnome', and the 'Hydrant' are composed of both shiny and rough parts. While the 'Duck', the 'Gnome', the green 'Vase', and the 'Sheep' comprise mostly homogeneous colors, both the 'Globe' and the 'Teapot' have very detailed textures. The 'Teapot' is painted, creating surface irregularities, while the 'Globe' is printed and, thus, smooth. For all objects, the number of bases $|B|$ is automatically determined by model selection within our framework. We visualize the more specular materials in red.*

4.2 Scene Representation

Let us assume a set of color images $\mathcal{I}_i : \mathbb{R}^2 \to \mathbb{R}^3$ captured from n different views $i \in \{1,\dots,n\}$. Without loss of generality, let us select $i = 1$ as the *reference view* based on which we will parameterize the surface geometry and materials as detailed below. Note that in the visualizations of this chapter, all observations are represented in this reference view.

Our goal is to jointly recover the camera poses, the scene's geometry, and the material properties in terms of svBRDFs. More formally, we wish to estimate the locations $\mathbf{x}_p = (x_p, y_p, z_p)^T$, surface normals $\mathbf{n}_p = (n_p^x, n_p^y, n_p^z)^T$, and svBRDFs $\rho_p(\cdot)$ of a set P of surface points $p \in P$, as well as the projective mappings $\pi_i : \mathbb{R}^3 \to \mathbb{R}^2$ that map a 3D point $\mathbf{x}_p \in \mathbb{R}^3$ into camera image i. We assume that each image is illuminated by exactly one point light source. Similar to prior works, we assume that global and ambient illumination effects are negligible.

We now describe the parameterizations of our model: Our camera representation in Sec. 4.2.1, the geometry representation through depth and normal maps in Sec. 4.2.2, and our material representation using svBRDFs in Sec. 4.2.3.

4.2.1 Camera Representation

We use a perspective pinhole camera model and assume constant intrinsic camera parameters that can be estimated using established calibration procedures [Zha99]. We also assume that all images have been undistorted and the vignetting has been removed. Therefore, we only optimize for the extrinsic parameters (i.e., rotation and translation) of each projective mapping $\pi_i : \mathbb{R}^3 \rightarrow \mathbb{R}^2$.

4.2.2 Geometry Representation

We parameterize geometry in terms of both depth and normal maps and enforce consistency between them using soft constraints (the regularizer is presented in the following Sec. 4.3.2).

Depth Map: We define all surface points P in terms of the depth map in the reference view $\mathcal{Z}_1 = \{z_p\}$, using $p \in P$ as the pixel/point index. Assuming a pinhole projection, the 3D location of surface point p is given by

$$\begin{aligned} \mathbf{x}_p &= \pi_1^{-1}(u_p, v_p, z_p) \\ &= \left(\frac{u_p - c_x}{f}, \frac{v_p - c_y}{f}, 1 \right) z_p \end{aligned} \tag{4.1}$$

where $[u_p, v_p]^T$ denotes the location of pixel p in the reference image $\mathcal{I}_c$, z_p is the depth at pixel p, π_1^{-1} is the inverse projection function and f, c_x, c_y denote its parameters.

Normal Map: We represent normals $\mathbf{n}_p$ as 3D unit vectors. In every iteration of the gradient-based optimization, we estimate an angular change for this vector so that we avoid both the unit normal constraint and the gimbal lock problem.

4.2.3 Material Representation

To model the reflectance properties of a material, we use a parametric version of the spatially-varying BRDF $\rho_p(\mathbf{n}_p, \omega_{\text{in}}, \omega_{\text{out}})$ as introduced in Sec. 2.1.4, and estimate its parameters per pixel/point p.

svBRDF: The svBRDF $\rho_p(\mathbf{n}_p, \omega_{\text{in}}, \omega_{\text{out}})$ models the fraction of light that is reflected from incoming light direction ω_{in} to outgoing light direction ω_{out} given the surface normal $\mathbf{n}_p$ and 3D location $\mathbf{x}_p$ at point $p \in P$. We use a modified version of the Cook-Torrance microfacet svBRDF model [CT82], which we presented in Eq. (2.25):

$$\rho_p(\mathbf{n}_p, \omega_{\text{in}}(\mathbf{x}_p), \omega_{\text{out}}(\mathbf{x}_p)) = \mathbf{d}_p + \mathbf{s}_p \frac{D(r_p)\, G(\mathbf{n}_p, \omega_{\text{in}}(\mathbf{x}_p), \omega_{\text{out}}(\mathbf{x}_p), r_p)}{4(\mathbf{n}_p \cdot \omega_{\text{in}}(\mathbf{x}_p))(\mathbf{n}_p \cdot \omega_{\text{out}}(\mathbf{x}_p))} \tag{4.2}$$

where $D(\cdot)$ describes the microfacet slope distribution, $G(\cdot)$ is the geometric attenuation factor, and $\mathbf{d}_p \in \mathbb{R}^3$, $\mathbf{s}_p \in \mathbb{R}^3$, and $r_p \in \mathbb{R}$ denote diffuse albedo, specular albedo, and surface roughness, respectively. We use Smith's function as implemented in Mitsuba [Jak10] for $G(\cdot)$, see Eq. (2.22) and Eq. (2.23), and the GTR model of the Disney svBRDF [Bur12] for

$D(\cdot)$, Eq. (2.21) for $\gamma = 1$. Following [Nam+18], we ignore the Fresnel effect, which cannot be observed using a handheld setup.

As illustrated in Fig. 4.2, many objects can be modeled well with few specular material components [LN16] while the object texture is more complex. We thus allow the diffuse albedo $\mathbf{d}_p$ to vary freely per pixel p, and model specular reflectance as a combination of $|B|$ specular base materials $b \in B$

$$\binom{\mathbf{s}_p}{r_p} = \sum_{b \in B} \alpha_p^b \binom{\mathbf{s}_b}{r_b} \tag{4.3}$$

with per-pixel svBRDF weights $\alpha_p^b \in [0,1]$ and specular base materials $\{(\mathbf{s}_b, r_b)\}_{b \in B}$. Note that this contrasts other representations [Len+03; Gol+10] which also linearly combine the diffuse part, hence requiring more base materials to reach the same fidelity. We found that $B \leq 3$ specular bases are sufficient for almost all objects. In summary, our svBRDF is fully determined by $\{(\mathbf{s}_b, r_b)\}_{b \in B}$ and $\{\mathbf{d}_p, \alpha_p\}_{p \in P}$.

4.3 Optimization Objective

This section describes our objective function. Let $\mathcal{X} = \{\{(z_p, \mathbf{n}_p, \rho_p)\}_{p \in P}, \{\pi_i\}_{i=2}^n\}$ denote the depth, normal and material for every pixel p in the reference view, as well as the projective mapping for each adjacent view. We formulate the following objective function

$$\mathcal{X}^* = \underset{\mathcal{X}}{\operatorname{argmin}} \; \psi_{\mathcal{P}} + \psi_{\mathcal{GC}} + \psi_{\mathcal{D}} + \psi_{\mathcal{N}} + \psi_{\mathcal{M}} \tag{4.4}$$

omitting the dependency on $\mathcal{X}$ and the relative weights between the individual terms. Our objective function is composed of five terms which encourage photo-consistency $\psi_{\mathcal{P}}$, geometric consistency $\psi_{\mathcal{GC}}$, depth compatibility $\psi_{\mathcal{D}}$, normal smoothness $\psi_{\mathcal{N}}$ and material smoothness $\psi_{\mathcal{M}}$. We elaborate on all terms in the following.

4.3.1 Photo-Consistency

The photo-consistency term ensures that the prediction of our model matches the observation $\mathcal{I}_i$ for every image i and pixel p:

$$\psi_{\mathcal{P}}(\mathcal{X}) = \frac{1}{n} \sum_i \sum_p \| \, \varphi_p^i \, [\mathcal{I}_i(\pi_i(\mathbf{x}_p)) - \mathcal{R}_i(\mathbf{x}_p, \mathbf{n}_p, \rho_p)] \, \|_1 \tag{4.5}$$

Here, $\mathcal{R}_i$ denotes the *rendering operator* for image i which applies the rendering equation [Kaj86] to every pixel p, recall Sec. 2.1.5. Assuming a single point light source, we obtain

$$\mathcal{R}_i(\mathbf{x}_p, \mathbf{n}_p, \rho_p) = \rho_p \left(\mathbf{n}_p, \omega_{\text{in}}^i(\mathbf{x}_p), \omega_{\text{out}}^i(\mathbf{x}_p)\right) \frac{a_i(\mathbf{x}_p) \, \mathbf{n}_p^T \, \omega_{\text{in}}^i(\mathbf{x}_p)}{d_i(\mathbf{x}_p)^2} \, L \tag{4.6}$$

where $\omega_{\text{in}}^i(\mathbf{x}_p)$ denotes the direction of the ray from the surface point $\mathbf{x}_p$ to the light source and $\omega_{\text{out}}^i(\mathbf{x}_p)$ denotes the direction from $\mathbf{x}_p$ to the camera center. $a_i(\mathbf{x}_p)$ is the angle-dependent light attenuation which is determined through photometric calibration, $d_i(\mathbf{x}_p)$ is the distance between $\mathbf{x}_p$ and the light source, and L denotes the radiant intensity of the light. Note that all terms depend on the image index i, as the location of the camera and the light source vary from frame to frame when recording with a handheld light stage.

The visibility term φ_p^i in (4.5) disables occluded or shadowed observations, i.e., we do not optimize for these regions. We set $\varphi_p^i = 1$ if surface point $\mathbf{x}_p$ is both visible in view i (i.e., no occluder between $\mathbf{x}_p$ and the i'th camera) *and* illuminated (i.e., no occluder between $\mathbf{x}_p$ and the point light), and $\varphi_p^i = 0$ otherwise. Note that for the reference view, every pixel is visible but not necessarily illuminated.

4.3.2 Geometry Regularization

Geometric Consistency: We enforce consistency between depth $\{z_p\}$ and normals $\{\mathbf{n}_p\}$ by ensuring that the normal field integrates to the estimated depth map. We formulate this constraint by maximizing the inner product between the estimated normals $\{\mathbf{n}_p\}$ and the

cross product of the surface tangents at $\{\mathbf{x}_p\}$:

$$\psi_{\mathcal{GC}}(\mathcal{X}) = -\sum_p \mathbf{n}_p^T \left(\frac{\frac{\partial z_p}{\partial x} \times \frac{\partial z_p}{\partial y}}{\| \frac{\partial z_p}{\partial x} \times \frac{\partial z_p}{\partial y} \|_2} \right) \tag{4.7}$$

As the derivations for surface tangents $\frac{\partial z_p}{\partial x}$ and $\frac{\partial z_p}{\partial y}$ are similar, we show only the derivation of the horizontal tangent:

$$\begin{aligned} \frac{\partial z_p}{\partial x} &\propto \left[1, 0, \frac{\partial \mathcal{Z}_1(\pi_1(\mathbf{x}_p))}{\partial x} \right]^T \\ &= \left[1, 0, \frac{\partial \mathcal{Z}_1(\pi_1(\mathbf{x}_p))}{\partial [u,v]^T} \frac{\partial [u,v]^T}{\partial x} \right]^T \\ &= \left[1, 0, \nabla \mathcal{Z}_1(\pi_1(\mathbf{x}_p))^T \, [f/z_p, 0]^T \right]^T \end{aligned} \tag{4.8}$$

where $\nabla \mathcal{Z}_1(\pi_1(\mathbf{x}_p))$ denotes the gradient of the depth map, which we estimate using finite differences.

A valid question to raise is whether a separate treatment of depth and normals is necessary. An alternative formulation would consider consistency between depth and normals as a hard constraint, i.e., enforcing Equation (4.7) strictly and optimizing only for depth. While reducing the number of parameters to be estimated, we found that such a representation is prone to local minima during optimization due to the complementary nature of the constraints (depth vs. normals/shading). Instead, using auxiliary normal variables and optimizing for both depth and normals using a soft coupling between them allows us to overcome these problems.

Depth Compatibility: The depth term allows for incorporating depth measurements $\mathcal{Z}_1$ in the reference view $i = 1$ by regularizing our estimates z_p against it:

$$\psi_{\mathcal{D}}(\mathcal{X}) = \sum_p \| z_p - \mathcal{Z}_1(u_p, v_p) \|_2^2 \tag{4.9}$$

Note that our model can significantly improve upon the initial coarse geometry provided by the structured light sensor by exploiting shading cues. However, as these cues are related to depth variations (i.e., normals) rather than absolute depth, they do not fully constrain the 3D shape of the object. Our experiments demonstrate that combining complementary depth and shading cues yields locally detailed and globally consistent reconstructions.

Normal Smoothness: We apply a standard smoothness regularizer to the normals of adjacent pixels $p \sim q$

$$\psi_{\mathcal{N}}(\mathcal{X}) = \sum_{p \sim q} \| \mathbf{n}_p - \mathbf{n}_q \|_2^2 \tag{4.10}$$

in order to encourage smooth surfaces.

4.3.3 Material Regularization

Material Smoothness: We observe specular svBRDF components only for a minority of pixels, the ones that actually observe a specular highlight in at least one of their measurements. Therefore, we introduce a non-local material regularizer that propagates specular behavior across image regions of similar appearance. Assuming that nearby pixels with similar diffuse behavior also exhibit similar specular behavior, we formulate this term by penalizing deviation of the material weights wrt. a bilaterally smoothed version of themselves

$$\begin{aligned} \psi_{\mathcal{M}}(\mathcal{X}) &= \sum_p \left\| \alpha_p - \frac{\sum_q \alpha_q w_q g_{pq}}{\sum_q w_q g_{pq}} \right\|_1 \\ &- \sum_p \left\| \alpha_p - \frac{1}{P} \sum_q \alpha_q \right\|_1 \end{aligned} \tag{4.11}$$

using a Gaussian kernel $g_{pq}(\mathbf{d}_p, \mathbf{d}_q)$ with 3D location $\mathbf{x}$ and diffuse albedo $\mathbf{d}$ at pixels p and q as features:

$$g_{pq} = \exp\left(-\frac{(\mathbf{x}_p - \mathbf{x}_q)_2^2}{2\sigma_1^2} - \frac{(\mathbf{d}_p - \mathbf{d}_q)_2^2}{2\sigma_2^2} \right) \tag{4.12}$$

As the most informative samples for specular material estimation are those that potentially observe a highlight, the weights $w_q = \max_i \text{acos}^{-1}(\mathbf{n}_q \cdot \mathbf{h}_q^i)$ with halfvector $\mathbf{h}_q^i$ (i.e., the bisector between ω_{in} and ω_{out}) and normal $\mathbf{n}_q$ increase the contribution of pixels q which are observed close to perfect mirror reflection in any view i. We use the permutohedral lattice [ABD10] to evaluate the bilateral filter efficiently.

The second term in (4.11) encourages material sparsity by maximizing the distance to the average svBRDF weights where P denotes the total number of surface points/pixels.

4.4 Optimization

We now describe the parameter initialization in Sec. 4.4.1, our model selection in Sec. 4.4.2, and how we minimize our objective function Eq. (4.4) with respect to $\mathcal{X}$ in Sec. 4.4.3.

4.4.1 Initialization

Initial Poses: The camera poses can be either initialized using classical *Structure from Motion (SfM)* pipelines such as COLMAP [SF16; Sch+16] or using a set of fiducial markers. As SfM approaches fail in the presence of textureless surfaces, we use a small set of AprilTags [WO16] attached to the table supporting the object of interest. As evidenced by our experiments, the poses estimated using fiducial markers are not accurate enough to model pixel-accurate light transport. We demonstrate that geometry and materials can be significantly improved by jointly refining the initial camera poses.

Initial Depth: The initial depth map $\mathcal{Z} = \{z_p\}$ can be obtained using active or passive stereo or the visual hull of the object. As we do not assume textured objects and silhouettes can be difficult to extract in the presence of dark materials, we use active stereo with a Kinect-like dot pattern projector for estimating $\mathcal{Z}$. More specifically, we estimate a depth map for each of the n views, integrate them using volumetric fusion [CL96], and project the resulting mesh back to the reference view.

Initial Normals and Albedo: Assuming a Lambertian scene, normals and albedo can be recovered in closed form. We follow the approach of Higo et al. [Hig+09] and use RANSAC to reject outliers due to specularities.

Initial Specular svBRDF Parameters and Weights: We initialize each pixel in the scene as a uniform mix of all base materials. To diversify the initial base materials, we initialize the specular base components $\mathbf{s}_b$ differently and set each base roughness r_b to 0.1.

4.4.2 Model Selection

We perform model selection by optimizing for multiple numbers of specular base materials $|B| \in \{1,2,3\}$, choosing that with the smallest photometric error while adding a small MDL penalty (linear in $|B|$).

4.4.3 Implementation

Due to memory constraints, we represent the inferred quantities $\mathcal{X}$ at half-resolution while utilizing the full-resolution observations. We implemented the rendering function $\mathcal{R}_i$ using PyTorch [Pas+17], exploiting PyTorch's GPU acceleration and auto-differentiation capabilities. We jointly optimize over the parameters $\mathcal{X}$ using ADAM [KB15] with hyperparameters $\beta_1 = 0.9$, $\beta_2 = 0.9$ and $\varepsilon = 10^{-3}$. We found these settings to be a good fit for our problem and ablate our choice in Sec. 4.6.2.

We choose the step size multipliers for each of the parameters in a physically-motivated way (because of the way ADAM works, this is a soft upper bound on the actual step size).

In our experiments, we empirically found the following values to perform well: 0.1mm for the depth, 0.5 degrees for normals, 0.1mm and 0.5 degrees, respectively, for poses, and 1% of the domain range for material properties.

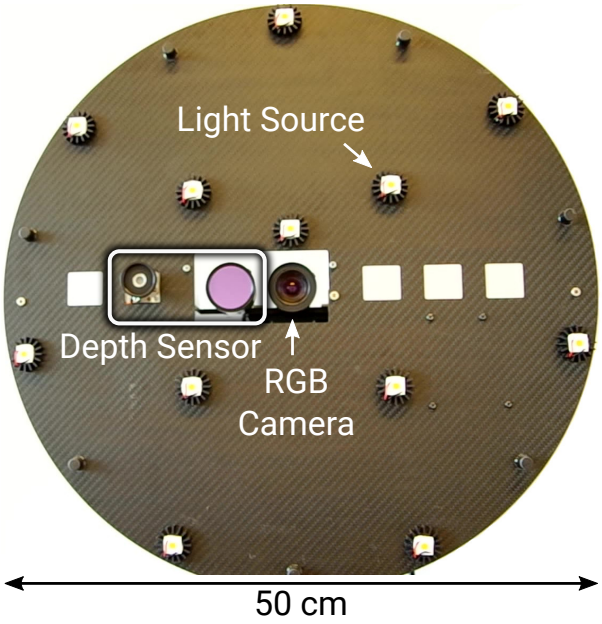


Figure 4.3: Sensor Rig. *Our custom-made handheld capture device features a high-resolution RGB camera, a Kinect-like active depth sensor, and* 12 *high-power LEDs (modeled as point light sources) that surround the camera in two circles (with radii* 10*cm and* 25*cm).*

4.5 Handheld Sensor

In order to evaluate our method quantitatively and qualitatively, we capture several real objects using a custom-built rig with active illumination. Reconstructions of these objects are shown in Fig. 4.2. We now describe our hardware in Sec. 4.5.1 and data capture procedure in Sec. 4.5.2.

4.5.1 Hardware

Our custom-built handheld sensor rig is shown in Fig. 4.3. While we use multiple light sources for a dense sampling of the BRDF, our framework and code are directly applicable to any number of lights.

Our custom-built handheld capture rig is powered by a compact battery pack such that it is easily portable. It comprises two high-resolution cameras, 12 high-powered LEDs, and a custom-built laser with a *Diffractive Optical Element (DOE)*.

The cameras have a native resolution of 4112×3008. The RGB camera uses a standard Bayer pattern (BGGR), while the *Infrared (IR)* camera is a single-channel camera with a high-pass filter that removes the visible spectrum. The cameras are hardware-synchronized with each other and the LEDs by an Arduino board. Empirically, we found an exposure time of 15ms to be a reasonable compromise between motion blur and image noise.

The laser emits light of 830nm at up to 650mW through a DOE that diffracts it into $\approx$

Observation Crop Reconstruction

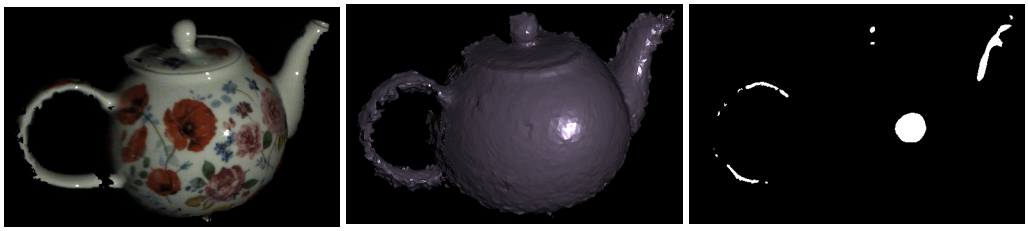
Figure 4.4: Super-Resolution and Denoising (2.5D). *Blurred and noisy input images (left and middle) get denoised and sharpened by our method (right).*

200000 separate beams in a pseudo-random pattern. Its projected dot pattern is subsequently observed by the IR camera; to keep the pattern and the wavelength constant, the laser module is actively cooled. We perform OpenCV block matching on the observed pattern to estimate a depth map for each of the observations. Those are then volumetrically fused into a consistent mesh using TSDF fusion [Zen+17]. We render the depth map $\mathcal{Z}$ from this mesh into the reference view and use this depth as the initial depth estimate.

The whole sensor rig is calibrated geometrically and photometrically in advance. For the geometric calibration, we determine the camera intrisics, response, and extrinsics as well as the positions and orientations of cameras, lights, and laser relative to each other. Photometrically, we calibrate the camera vignetting as well as the radiant intensities and angular attenuation curve per color channel of the light sources. In our method, we assume that debayering, undistorting, and devignetting of the input images is done in pre-processing.

4.5.2 Data Capture

The objects are placed on a table with AprilTags [WO16] for tracking the sensor position. The room is darkened completely, and we assume all ambient light to be negligible. We capture videos of each object by moving the handheld sensor around the object slowly. Simultaneously, we alternate the illumination such that each image is illuminated by exactly one light source. Given each reference view, we select 45 views within a viewing cone of 30 degrees by maximizing the minimum pairwise distance; no two views are ever close together. These views are then split into 40 training and five held-out test views. While a handheld setup is challenging due to the trade-off between motion blur and image noise, our experiments demonstrate that our method is capable to super-resolve and denoise fine textures while simultaneously rejecting blurry observations, see Fig. 4.4.



Observation | Initial Geometry | Specularity Map

Figure 4.5: Specularity Mask (2.5D). *We smooth the initial depth map and calculate the angle between normal* $\{\mathbf{n}_p\}$ *and halfvector for every pixel. We determine the in-specular and non-specular image regions by thresholding the resulting angles.*

4.6 Experimental Evaluation

In order to evaluate our method quantitatively and qualitatively, we capture several real objects using the custom-built rig with active illumination presented in Sec. 4.5. Reconstructions of these objects are shown in Fig. 4.2. We scanned the objects with an Artec Spider[1] to obtain ground truth geometry.

In this section, we first introduce the evaluation metrics, both geometric and photometric, in Sec. 4.6.1, and then provide an ablation study in terms of the individual components of our model in Sec. 4.6.2. Next, we compare our approach with several competitive baselines in Sec. 4.6.3 and present qualitative results on nine diverse objects in Sec. 4.6.4. Finally, we discuss limitations and give an outlook in Sec. 4.6.5.

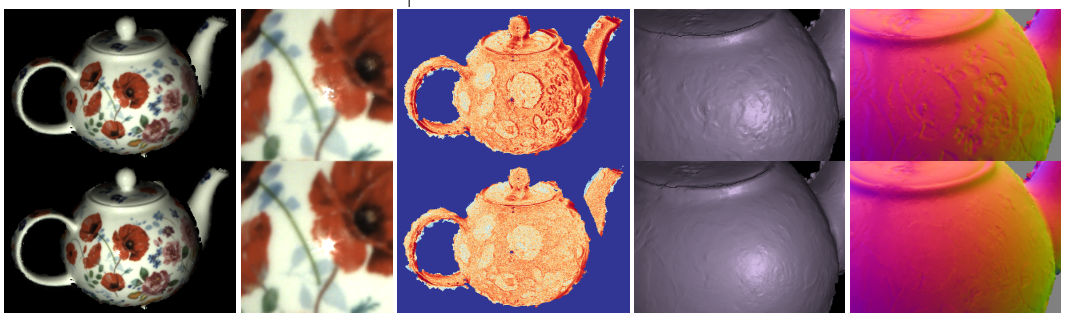
4.6.1 Evaluation Protocol

We evaluate the estimated shape $\{\mathbf{x}_p\}$ wrt. the Artec Spider scan. The ground truth scan is first roughly aligned by hand and subsequently finetuned using dense image-based alignment wrt. depth and normal errors. We evaluate geometric accuracy using the average point-to-mesh distance for all reconstructed points as in [Jen+14]. To evaluate surface normals, we calculate the average angular error (AAE) between the predicted normal $\mathbf{n}_p$ and the normal of the closest point in the ground truth scan.

To quantify photometric reconstruction quality, we calculate the photo-consistency term in Eq. (4.5) for the test views. It is given as the mean L_1 loss between the observation and our prediction for every pixel and all input images. Since most of the BRDF information is captured in pixels that potentially observe specular highlights, we split the mean photometric test error into pixels in possibly-specular and non-specular regions using a mask as the one illustrated in Fig. 4.5. To calculate this *specularity mask* for each observation, we threshold the angle between the normalized halfvector ($\propto \omega_{\text{in}} + \omega_{\text{out}}$) and the initial normal for each pixel and each observation. For this, the initial normals are calculated from the input depth and smoothed. The angle threshold is set to 15°. We empirically found that the resulting regions cover the majority of pixels in specular highlight areas.

[1] https://www.artec3d.com/portable-3d-scanners/artec-spider

Photometric Test Error	Overall	Specular	Non-Specular
Fixed Poses	1.210	3.349	1.151
Full Model	**1.138**	**3.243**	**1.081**



Estimation Crop Error Shape Normals

Figure 4.6: Pose Optimization (2.5D). *Compared to using the input poses (top), optimizing the poses (bottom) improves reconstructions, both quantitatively and qualitatively. The photometric error is reported for regions with and without specular highlights. See Fig. 4.5 for more details.*

4.6.2 Ablation Study

In this section, we demonstrate the need to optimize the camera poses and discuss the effect of the material smoothness and the geometric consistency terms. We also investigate the impact of the number of views on the photometric and geometric error. Finally, we ablate the influence of the input geometry on the final reconstruction quality.

Pose Optimization: Disambiguating geometric properties from the material is a major challenge. We found that optimizing the poses jointly with the other parameters is crucial for this, particularly when working with a handheld scanner. Fig. 4.6 shows that inaccurate poses cause significant contamination of the geometry with texture information. This is even more crucial when estimating specularities: Misalignment causes highlights to be inconsistent with the geometry and, therefore, difficult to recover.

Material Segmentation: Decomposing the object's appearance into its individual materials is an integral element of our approach. Our material smoothness term Eq. (4.11) propagates material information over large areas of the image. This is essential as we only obtain sparse BRDF measurements at each pixel. It leads to semantically meaningful segmentations, as illustrated in Fig. 4.2, and a more successful generalization, as shown in Fig. 4.7a.

Geometric Consistency: Splitting the depth and normals into separate optimization variables yields a better-behaved optimization problem, but coupling depth and normals, Eq. (4.7), proves crucial for consistent results. Even though the photometric term provides some constraints for the depth at each pixel, Fig. 4.7b shows that omitting the geometric consistency term results in high-frequency structure artifacts.

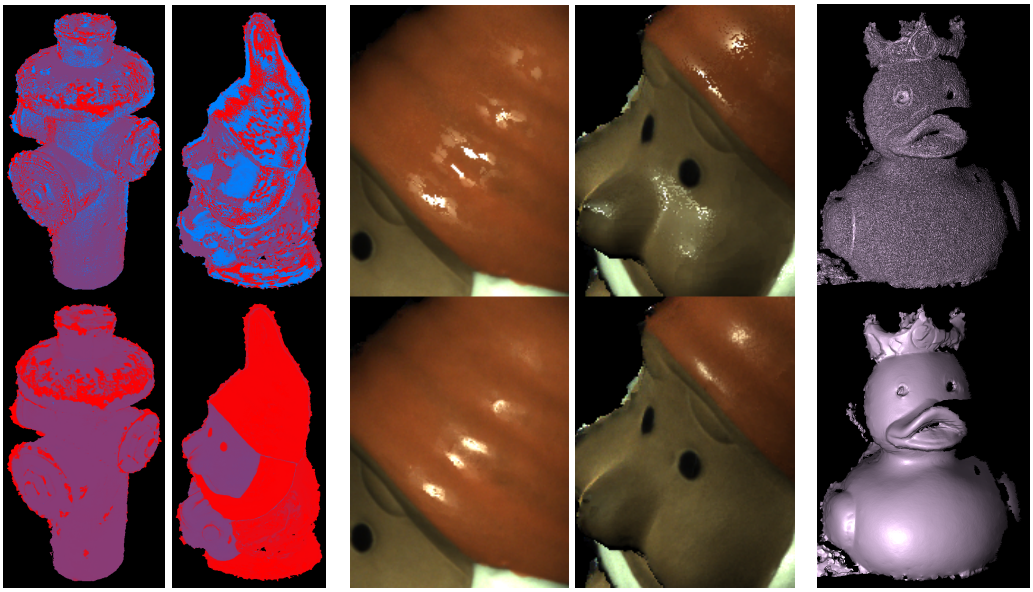


Figure 4.7: Loss Regularizers (2.5D). *Without the regularization (top), the appearance is inconsistent within homogeneous areas of the object. Using the regularization losses (bottom), we are able to propagate the information and successfully generalize to new illumination conditions on the test set.*

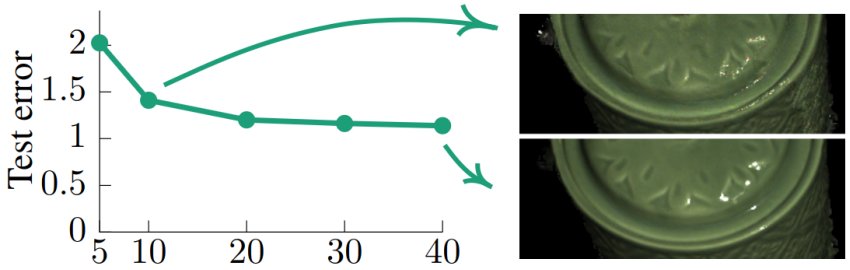
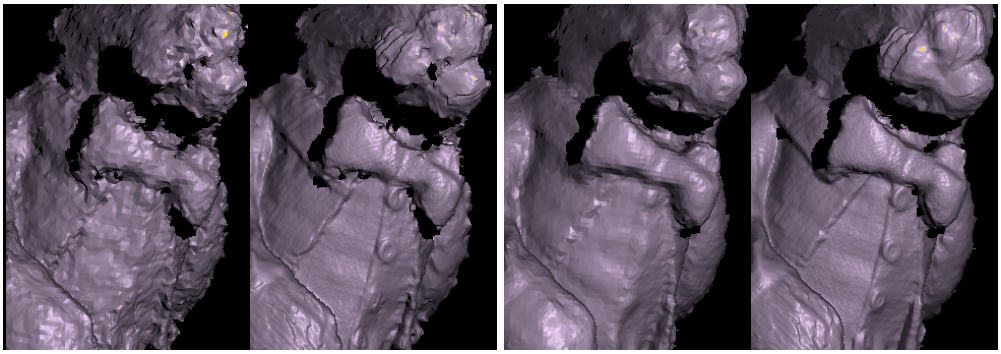
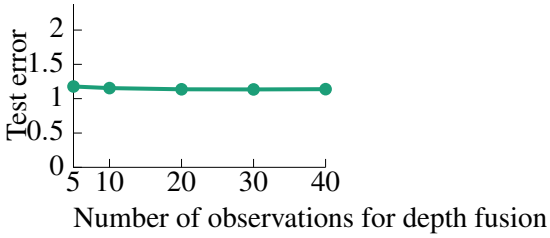


Figure 4.8: Number of Input Views (2.5D). *The photometric test error (green) degrades gracefully with decreasing number of observations. As expected, the quality of the highlights is most affected by a small number of views.*



Initialized from five depth maps Initialized from 40 depth maps

(a) Qualitative Comparison



(b) Quantitative Comparison

Figure 4.9: Geometry Refinement (2.5D). *The reconstruction quality of the final geometry estimate is barely affected by the input depth estimate. The test photometric error only degrades slightly when using the worst depth initialization, and details are recovered, even from a very coarse initialization.*

Number of Input Views: We aim to estimate the spatially-varying BRDF, but we only observe a very sparse set of samples for each surface point p. Reducing the number of images exacerbates this problem, as shown in Fig. 4.8. We see that our method degrades gracefully, with reasonable results even for using only 10 input images.

We also evaluate the robustness of our method wrt. the initial geometry by reducing the number of depth maps fused for initialization. As Fig. 4.9a shows, our method is able to recover from inaccurate depth initialization and achieves similar quality reconstructions even when initializing from only five depth maps. The photometric loss supports these observations quantitatively in Fig. 4.9b: The test error is almost constant over the number of input depth images. By construction, our model does not recover geometry absent in the initial estimate.

Optimizer: We quantitatively compare four different optimizer choices in Tab. 4.1:

1. Gradient descent with a fixed step size for each parameter that we optimize over,
2. L-BFGS,
3. ADAM with the default hyperparameters $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\varepsilon = 10^{-8}$, and
4. ADAM with our tuned hyperparameters $\beta_1 = 0.9$, $\beta_2 = 0.9$ and $\varepsilon = 10^{-3}$.

Photometric Test Error	Overall	Specular	Non-Specular
GD with fixed step size	1.159	**2.934**	1.110
L-BFGS	2.296	4.432	2.235
Adam, default parameters	1.148	3.186	1.092
Adam, our parameters	**1.131**	3.039	**1.078**

Table 4.1: Different Optimizers. *For a detailed discussion, please refer to the main text.*

Gradient descent struggles with the balance between bright and dark pixels. It optimizes and reconstructs bright pixels well, but the signal-to-noise ratio in dark pixels degrades performance in those areas drastically. L-BFGS is a second-order optimization algorithm that relies on a series of line searches. We found that this fails to optimize our objective function properly. We assume this is due to the strong inter-dependencies between the different variables, where the relatively low number of line searches that L-BFGS is able to perform in the same time budget as the other optimizers is insufficient. The default parameters for ADAM result in the well-known problem of strong perturbations around local minima, and for this reason, we increased ε to 10^{-3}, mitigating this issue. Finally, given the relatively low number of iterations (1000), the shorter memory $\beta_2 = 0.9$ helps ADAM to adapt more quickly.

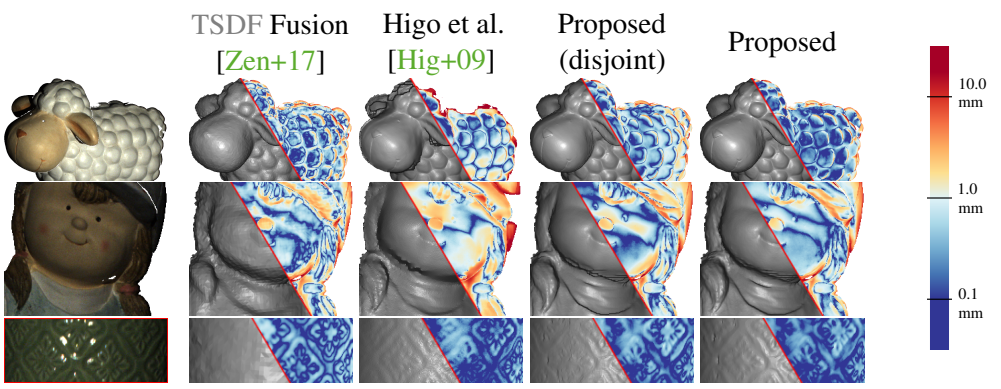


Figure 4.10: Qualitative Geometry Comparison (2.5D). *For each object, we show the rendered depth map (shaded based on estimated surface normals rather than the estimated normals $\mathbf{n}_p$) and the color-coded depth error wrt. the Artec Spider ground truth. We observe that the photometric approaches recover far more details than the purely geometric TSDF fusion. The resulting structure for Higo et al. [Hig+09] is rather noisy, while the disjoint version of our proposed approach is not as successful at disambiguating texture and geometry for very fine details. We refer to Fig. 4.11 and Fig. 4.12 for additional results.*

4.6.3 Comparison to Existing Approaches

Implementation of Existing Approaches: Similar to us, Higo et al. [Hig+09] use a handheld scanner to estimate depth, normals, and material using a 2.5D representation. Unlike us, they treat specular highlights, shadows, and occlusions as outliers using RANSAC. Georgoulis et al. [GPG14] and Nam et al. [Nam+18] also estimate structure and normals, explicitly modeling non-Lambertian materials. However, due to the nature of their pipelines, they are restricted to a disjoint optimization procedure and update geometry and materials in alternation. It is important to note that the baselines expect the camera positions to be known accurately. So for the baselines, we first refine the poses using SfM [Sch+16].

Unfortunately, none of the existing works provide code. We have re-implemented the approach of Higo et al. [Hig+09] as a baseline. To investigate the benefits of joint optimization, we implemented a *disjoint* variant of our method that alternates between geometry and material updates. We also evaluate the improvement over the initial TSDF fusion using the implementation of Zeng et al. [Zen+17] and compare against the MVS baseline COLMAP [SF16].

Qualitative and Quantitative Experiments: Our experimental evaluation is shown in Fig. 4.10 - Fig. 4.12 and Tab. 4.2. Additionally, we extend Tab. 4.2 by showing all its quantitative results separately per reference view and including the COLMAP [SF16] MVS baseline in Tab. 4.3.

COLMAP assumes, with some robustness, appearance constancy – this assumption is very much violated in our set-up with glossy materials and textureless objects. This becomes obvious for objects like the 'Duck', which has both shiny materials showing specular highlights and large textureless areas for which shading effects primarily dominate

		Duck	Vase	Girl	Gnome	Sheep	Hydrant	Rabbit
AEA	TSDF Fusion [Zen+17]	0.81	1.24	1.11	0.73	0.79	1.35	**2.16**
	Higo et al. [Hig+09]	2.65	1.05	1.59	1.60	2.09	1.81	2.85
	Proposed (disjoint)	0.81	**1.00**	1.06	0.65	0.64	1.18	2.25
	Proposed	**0.80**	**1.00**	**1.00**	**0.64**	**0.57**	**1.16**	**2.16**
AAE	TSDF Fusion [Zen+17]	6.75	12.09	11.40	7.64	8.38	11.67	24.42
	Higo et al. [Hig+09]	7.77	10.62	11.30	9.24	8.65	15.27	27.67
	Proposed (disjoint)	6.01	9.13	9.81	6.41	6.66	9.59	**23.01**
	Proposed	**5.17**	**8.98**	**8.73**	**5.74**	**5.60**	**8.50**	23.50

Table 4.2: Quantitative Geometry Comparison Overview (2.5D). *We report both the average Euclidean accuracy and average angular error, as discussed in Sec. 4.6.1. Please refer to Tab. 4.3 for a more detailed table.*

		Duck		Vase		Girl		Gnome		Sheep		Hydrant		Rabbit	
Metric	Method	View 1	View 2	View 1	View 2	View 1	View 2	View 1	View 2	View 1	View 2	View 1	View 2	View 1	View 2
AEA	COLMAP refined	4.96	4.80	1.78	0.44	1.62	1.62	4.01	2.84	3.11	2.12	**0.93**	1.14	**2.42**	1.94
	TDSF fusion (40 frames)	0.71	**0.91**	1.96	0.52	1.28	0.95	0.84	0.63	0.97	0.60	1.49	1.22	2.67	**1.66**
	Higo COLMAP poses	2.65	2.65	1.65	0.44	1.61	1.57	1.81	1.39	2.98	1.20	1.88	1.74	3.01	2.69
	Disjoint	0.67	0.96	**1.54**	0.47	1.23	0.88	0.75	0.56	0.74	0.54	1.25	1.12	2.75	1.75
	Proposed	**0.60**	0.99	1.59	**0.41**	**1.20**	**0.80**	**0.73**	**0.55**	**0.71**	**0.43**	1.24	**1.08**	2.63	1.68
AAE	COLMAP refined	39.56	56.19	12.93	12.72	24.46	23.27	42.78	38.85	38.80	30.97	20.31	18.95	27.48	24.64
	TDSF fusion (40 frames)	6.41	7.09	14.50	9.67	11.61	11.19	7.84	7.44	9.29	7.47	12.90	10.44	25.90	22.94
	Higo COLMAP poses	7.79	7.75	11.54	9.69	10.76	11.84	9.22	9.26	10.05	7.26	14.66	15.87	29.59	25.75
	Disjoint	6.24	5.77	**9.95**	8.31	10.13	9.49	6.54	6.29	6.89	6.42	9.26	9.92	**24.77**	**21.25**
	Proposed	**5.50**	**4.85**	10.53	**7.42**	**9.03**	**8.43**	**6.15**	**5.34**	**6.15**	**5.05**	**9.02**	**7.99**	25.15	21.86

Table 4.3: Quantitative Geometry Comparison (2.5D). *A more detailed version of Tab. 4.2. We report both the average Euclidean accuracy and average angular error.*

appearance. For more textured objects or less glossy materials, such as the 'Rabbit' or the 'Girl', COLMAP produces reasonable results.

We see that TSDF fusion, a purely geometric approach, reconstructs the general surface well but misses fine details. Their spatial regularizer helps Higo et al. [Hig+09] to achieve reasonable reconstructions, which are, however, strongly affected by the noisy, unregularized, normal estimates. Additionally, the RANSAC approach to shadow handling results in artifacts around depth discontinuities. Both the *joint* and *disjoint* versions of our approach reconstruct the scene more accurately than the baselines, but the joint approach consistently obtains better reconstruction accuracy given a fixed computational budget.

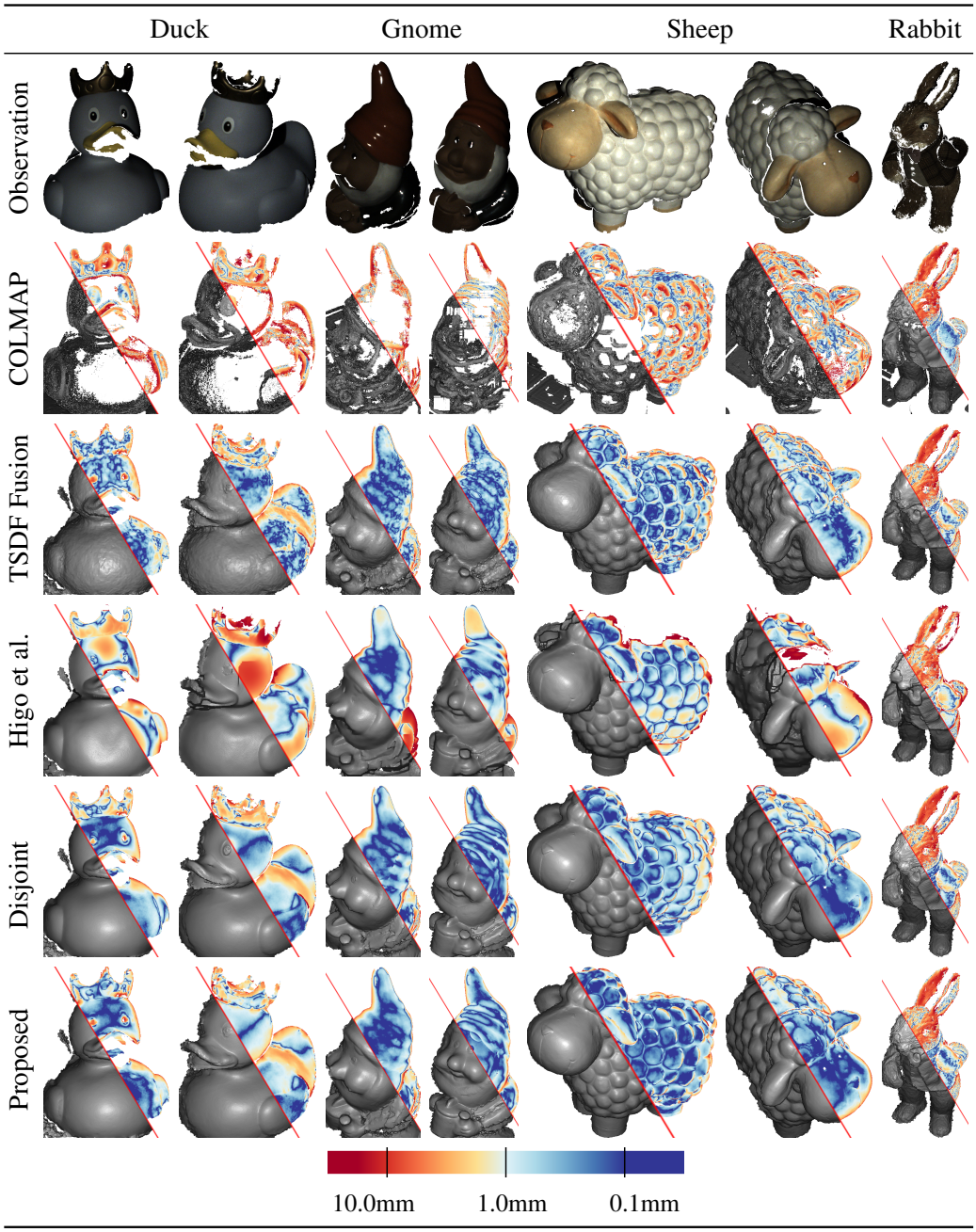


Figure 4.11: Qualitative Geometry Comparison (2.5D) *for the objects 'Duck', 'Gnome', 'Sheep', and 'Rabbit'. We compare COLMAP [SF16], TSDF Fusion [Zen+17], Higo et al. [Hig+09], a disjoint variant of our method, and our proposed method. For each method, object, and reference view, we show the rendered depth map (shaded based on the estimated surface normals) and the color-coded depth error wrt. the Artec Spider ground truth.*

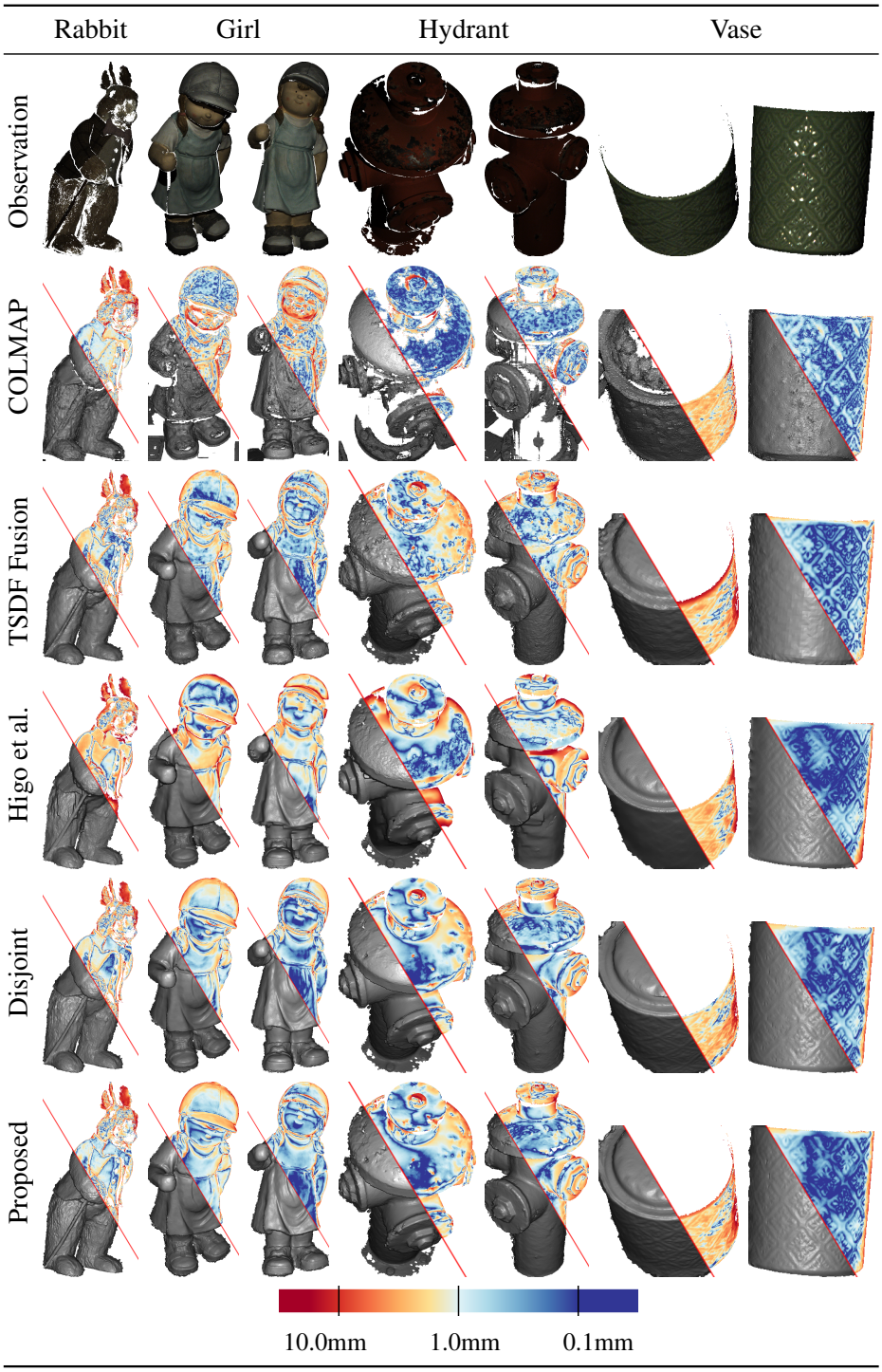


Figure 4.12: Qualitative Geometry Comparison (2.5D) *for the objects 'Rabbit', 'Girl', 'Hydrant', and 'Vase'. We compare COLMAP [SF16], TSDF Fusion [Zen+17], Higo et al. [Hig+09], a disjoint variant of our method, and our proposed method. For each method, object, and reference view, we show the rendered depth map (shaded based on the estimated surface normals) and the color-coded depth error wrt. the Artec Spider ground truth.*

4.6.4 Reconstruction Results

We show the reconstructions of our method for nine objects qualitatively in Fig. 4.13 - Fig. 4.15. Here, we highlight some observations:

- Details like the mouth and nose of the 'Sheep', the stitches on the cap of the 'Girl', and the rough woolen fabric of the 'Rabbit's' jacket are part of the statue's geometries. These fine structures (on the order of 1mm) get recovered successfully by the proposed method.
- As the 'Globe' and the 'Teapot' display highly detailed textures, optimization over camera poses is particularly important for these objects. Without it, the texture estimate is blurry as correspondences are offset. In contrast, the results show sharp textures, revealing well-aligned poses.
- The proposed method recovers most of the specular highlights on the body of the 'Vase' – this is only possible because of the detailed reconstruction of the geometry and normals.
- The 'Hydrant' or the eyes of the 'Duck' consists of two different materials, where one is more specular than the other. The segmentation images show that our proposed method cleanly separates these materials.
- The 'Gnome' and the 'Duck' combine materials with significantly different specular properties, which is well captured by the segmentation. Yet, for one reference view, the beard of the 'Gnome' is assigned wrong due to too few photometric cues. More observations showing specularities would resolve this issue.
- A lacquer finish lies on top of the delicate artwork of the 'Teapot'. Therefore, a single specular base material is sufficient to model its specular behavior.
- Contrary to previous objects, there is a smooth transition between colors on the dress of the 'Girl'. The per-pixel albedo estimation manages to recover that accurately.
- For the 'Globe', a strong texture is paired with a very smooth and shiny surface, making the disentanglement of geometry and appearance very challenging. The texture is well captured in the albedo map, yet some details are also present in the normal map. Even though the proposed approach is generally able to distinguish texture from geometry, there is still some ambiguity left; more input observations would help to disambiguate these further.
- There is no initial depth from our structured light sensor for the eyes of the 'Rabbit' and the small golden pineapple on the 'Vase'. Our method does not fill in missing geometry, so these parts are not modeled in the reconstructions.

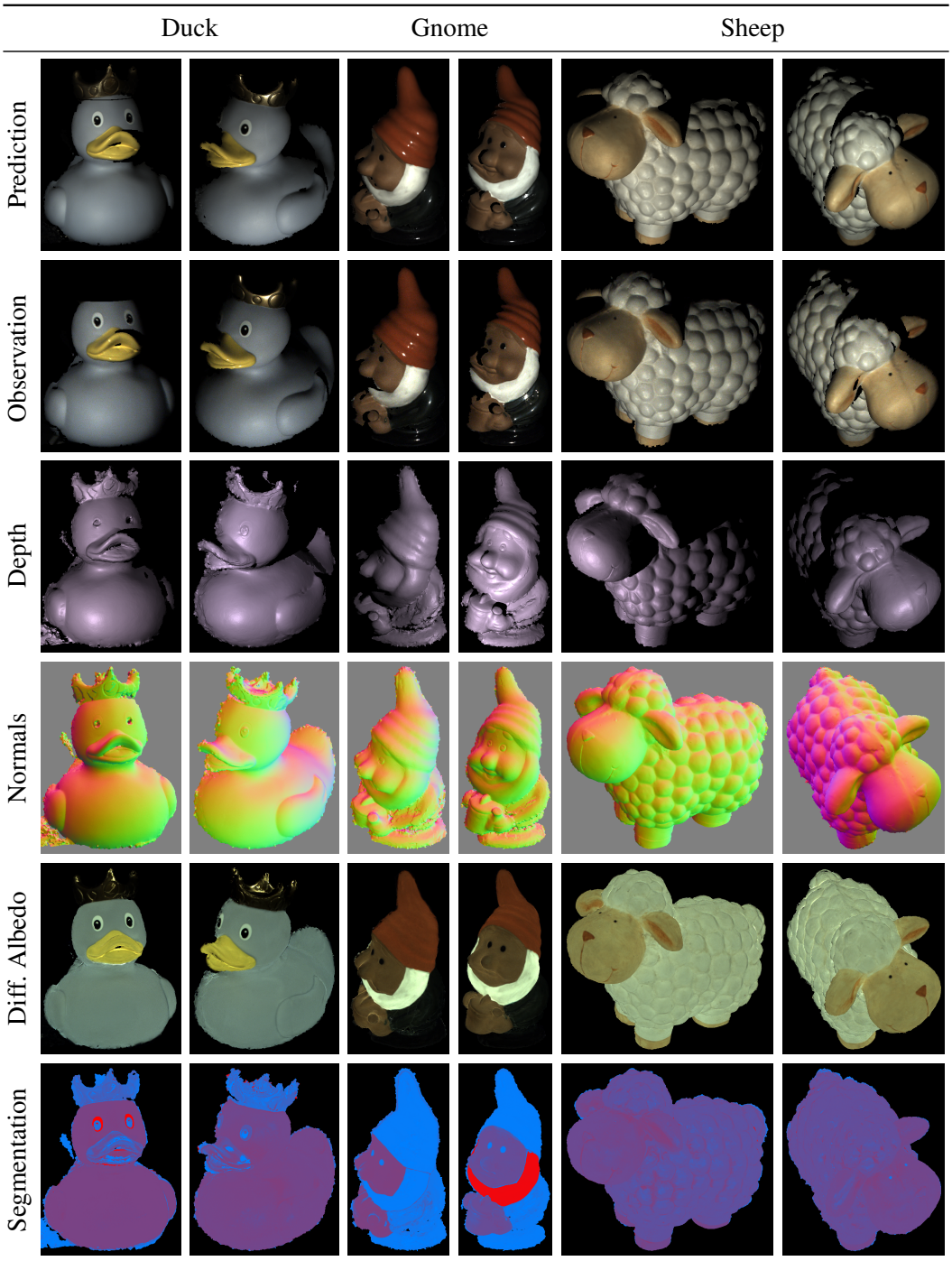


Figure 4.13: Reconstruction Results (2.5D) *for the objects 'Duck', 'Gnome', and 'Sheep'. Per object, we chose two reference views and a randomly selected observation each. We show the rendered prediction by our proposed method, the reconstructed geometry, normal, diffuse albedo, and specular material weight map. The normals used for rendering the geometry map are calculated from the predicted depth map.*

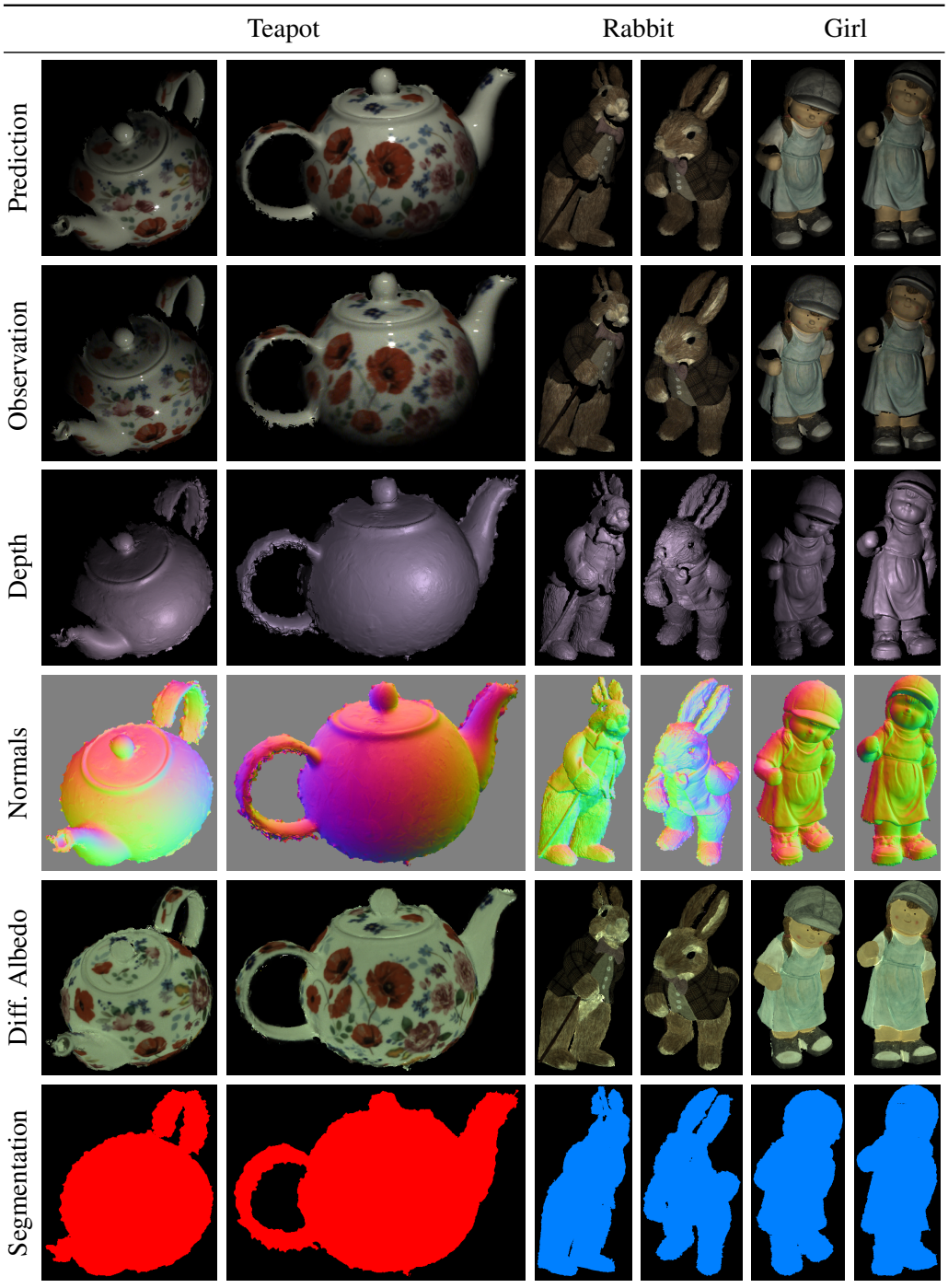


Figure 4.14: Reconstruction Results (2.5D) *for the objects 'Teapot', 'Rabbit', and 'Girl'. Per object, we chose two reference views and a randomly selected observation each. We show the rendered prediction by our proposed method, the reconstructed geometry, normal, diffuse albedo, and specular material weight map. The normals used for rendering the geometry map are calculated from the predicted depth map.*

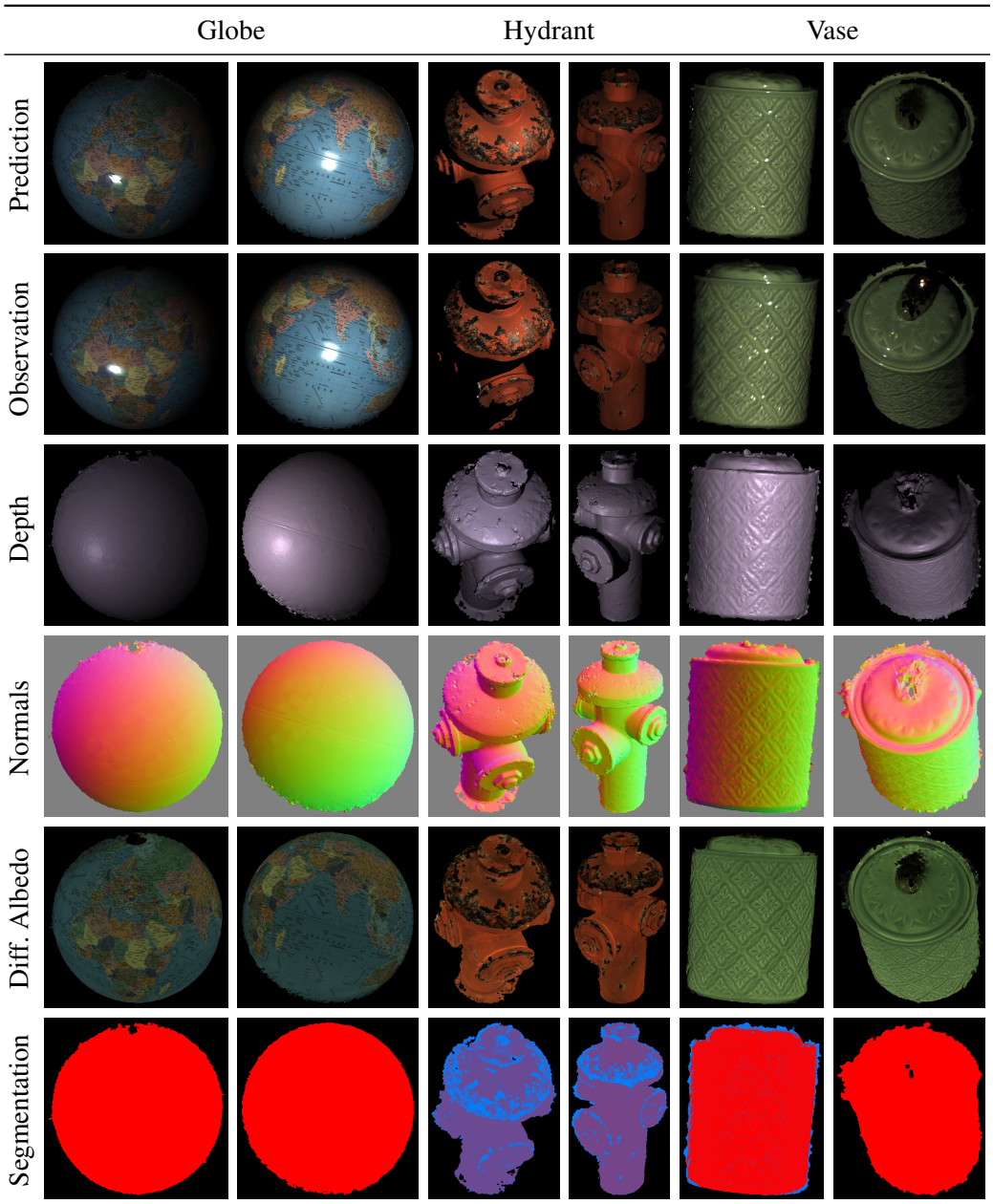


Figure 4.15: Reconstruction Results (2.5D) *for the objects 'Globe', 'Hydrant', and 'Vase'. Per object, we chose two reference views and a randomly selected observation each. We show the rendered prediction by our proposed method, the reconstructed geometry, normal, diffuse albedo, and specular material weight map. The normals used for rendering the geometry map are calculated from the predicted depth map.*

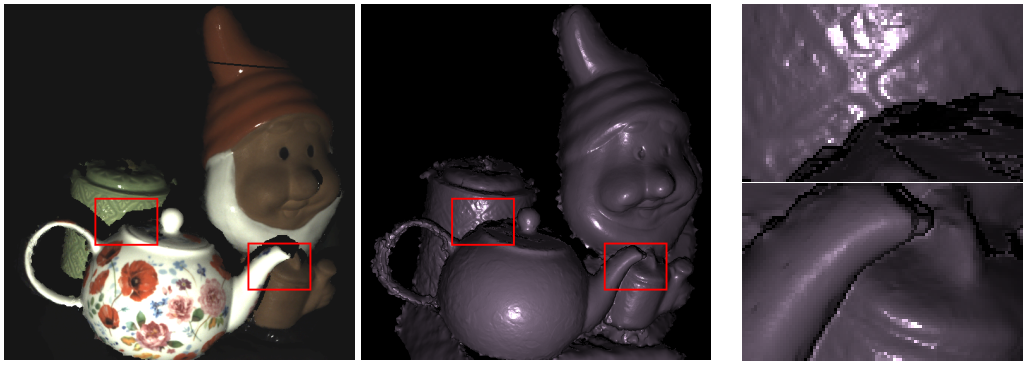


Figure 4.16: A strongly Non-Convex Scene (2.5D). *Due to our explicit shadow and occlusion model, our approach is also applicable to the reconstruction of more complex scenes.*

4.6.5 Limitations and Outlook

We have proposed a practical approach to estimating geometry and materials from a handheld sensor in 2.5D. By optimizing jointly over all parameters, including camera poses, our method is able to recover accurate geometry and material properties and produce semantically meaningful material weights, despite the complex correlation of geometry and materials. Additionally, the material smoothness term proved to be an effective regularizer for handling sparse material sample data from handheld capture systems. Finally, Fig. 4.16 illustrates that our approach is also able to handle strongly non-convex scenes of multiple objects whose shadows and occlusions often cause issues for existing methods.

Glossy black materials, such as the eyes of the 'Duck' (Fig. 4.7b) or the 'Rabbit' (Fig. 4.2), remain a significant challenge. For such materials, the signal-to-noise of the diffuse component is low, and the signal from specular highlights is very sparse so that neither the photo-consistency nor the depth compatibility term constrains the solution correctly.

Further, we show in our follow-up work that obtaining a full 3D model from multiple 2.5D reconstructions of the above method is problematic. Independently optimized 2.5D representations of a given object or scene do not align well and show many inconsistencies in the reconstructed parameter maps. We present these results in the upcoming Chap. 5 (Fig. 5.10) and propose a solution: By introducing a block-based optimization scheme that promotes global multi-view consistency, we obtain accurate and full 3D reconstructions. For it, the processing memory is independent of the scene size, allowing for reconstructions of larger scenes of multiple objects with many observation views.

5 Towards Scalable Multi-View Reconstruction of Geometry and Materials

This dissertation targets photo-realistic 3D reconstructions of multi-object scenes. Towards this goal, we have presented a solution for partial 2.5D reconstructions of objects in the previous chapter: Using joint optimization and effective regularization, the method in Chap. 4 accurately recovers geometry, material parameters, and camera poses of arbitrary objects given data captured by a handheld sensor system.

In this chapter, we extend the 2.5D method to full 3D reconstructions of scenes that exceed object-scale. We first analyze the previous method and find that a simple fusion of the resulting 2.5D parameter maps is insufficient to obtain an accurate and consistent 3D model. Consistency between keyframes does not emerge automatically without regularization despite partially overlapping fields of view. Additionally, targeting scenes of multiple objects requires a scalable scene representation and an efficient optimization strategy to stay within processing memory limits.

Therefore, we greatly extend the previous 2.5D method: We facilitate scalability to large numbers of observation views and optimization variables by introducing a distributed optimization algorithm that reconstructs 2.5D keyframe-based representations of the scene. Additionally, a novel multi-view consistency regularizer effectively synchronizes neighboring keyframes such that the local optimization results allow for seamless integration into a globally consistent 3D model.

This work is accepted for publication in TPAMI 2023, [Sch+23].

The chapter is organized as follows: First, we give an in-depth overview of the proposed method in Sec. 5.1. We then describe our method: We introduce our scene representation and parameterizations of the optimization variables in Sec. 5.2, present the model formulation and optimization objective in Sec. 5.3, and discuss the multi-view consistent optimization scheme in Sec. 5.4. Then, the mesh generation step that integrates the 2.5D optimization results into a full 3D model is explained in Sec. 5.5. In Sec. 5.6, we evaluate our method experimentally: We provide a study on the importance of each component in our formulation and show that our method compares favorably to baselines. We further demonstrate that our method accurately reconstructs various objects and materials and allows for expansion to spatially larger scenes. We believe this work represents a significant step towards making geometry and material estimation from handheld scanners scalable.

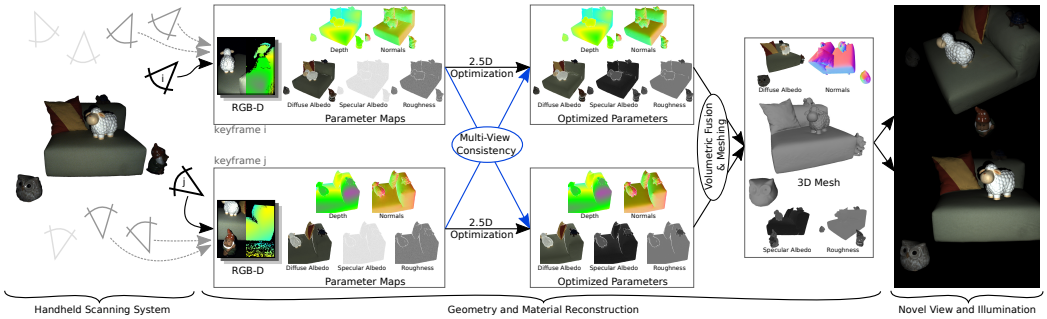


Figure 5.1: Globally Consistent Material and Geometry Reconstruction. *Given RGB-D images from a mobile handheld scanner (left), the proposed method uses local 2.5D representations to reconstruct globally consistent poses, geometry, and material parameter maps that allow for integration into a 3D representation with per voxel normals and material parameters (middle). This allows for rendering novel views under unseen illumination (right). Our approach can handle both multi-object scenes and very specular materials.*

5.1 Introduction

Reconstructing large scenes captured from many viewpoints at high resolution (e.g., 4K) becomes intractable quickly. In order to process larger multi-object scenes, a scalable scene representation is mandatory. Therefore, we propose to use local 2.5D scene representations and an optimization scheme that encourages global consistency between them. By optimizing in 2.5D, the proposed model has a constant memory footprint independent of the scene size and allows for reconstructing geometry and materials at larger scales, see Fig. 5.1. We summarize our contributions as follows:

- We extend our 2.5D method of Chap. 4 by a distributed optimization scheme over a set of 2.5D scene representations, enabling accurate integration of these reconstructions into full 3D models. We show that despite overlapping fields of view, regularizing multi-view consistency is crucial for globally accurate reconstructions without visual artifacts.
- We provide a study on the importance of each component in our formulation and a comparison to multiple baselines, including [Sch+20] and [Nam+18].
- We demonstrate that our model can be used to reconstruct scenes exceeding object-level at a high resolution which include multiple objects with various different materials.
- We provide videos of our reconstructed models and make our source code and dataset publicly available at https://sites.google.com/view/material-fusion/.

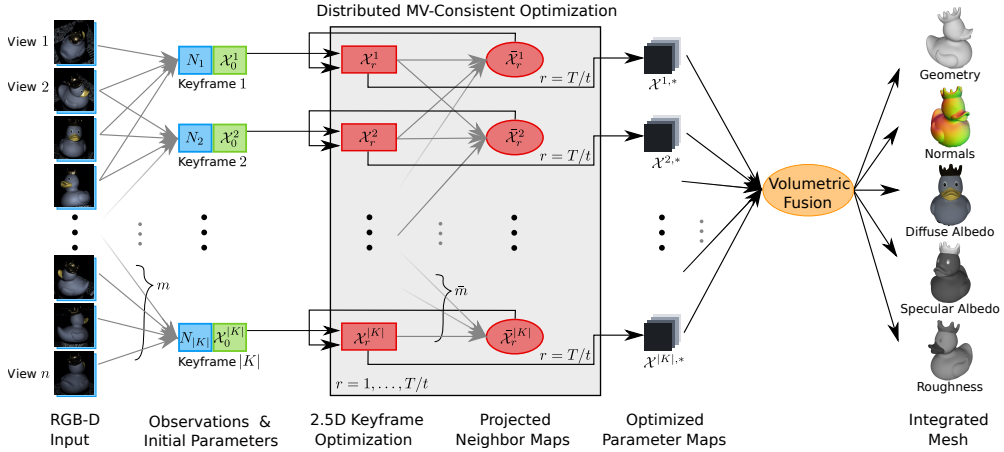


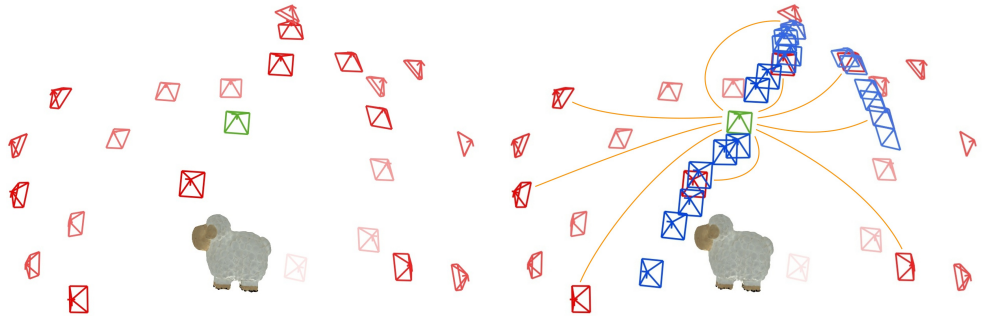
Figure 5.2: Pipeline Overview. *The input to our model is n RGB-D images from which we select a subset of well-distributed keyframes K. For each keyframe $k \in K$, we select m neighboring observation views N_k and initialize the set of optimization parameters $\mathcal{X}_0^k$. During optimization, we then iterate the following rounds r (gray box): After optimizing each keyframe representation independently for t iterations, we project the current parameter maps of all neighboring keyframes $\bar{N}_k$ into each keyframe k and use the resulting set $\bar{\mathcal{X}}_r^k$ as an additional constraint to the optimization of the next round $r+1$. The sets of $\bar{m}$ neighboring keyframes $\{\bar{N}_k\}_k$ are defined at the start for all keyframes k. After $r = T/t$ rounds, the resulting sets of optimized 2.5D parameter maps $\{\mathcal{X}^{k,*}\}_k$ are integrated into a full 3D model, represented by a mesh with per-vertex normal, diffuse and specular albedo, as well as roughness parameters.*

5.2 Scene Representation

Our goal is to reconstruct geometry, material properties, and camera poses from RGB-D data. Unfortunately, representing an entire scene in memory is computationally demanding, particularly when using memory-limited but computationally efficient GPUs for optimization. Towards scalable scene reconstruction, we, therefore, exploit a keyframe-based 2.5D representation that locally describes and optimizes geometry, materials, and poses. In particular, we adopt alternating block coordinate optimization of keyframes to minimize photometric errors while encouraging consistency between adjacent keyframes using soft constraints. An overview of our method is shown in Fig. 5.2.

The input to our model is an RGB-D sequence captured with a handheld scanner, as shown in Fig. 4.3, that consists of a color image $\mathcal{I}_i : \mathbb{R}^2 \to \mathbb{R}^3$ and a depth map $\mathcal{Z}_i : \mathbb{R}^2 \to \mathbb{R}$ at each frame $i \in N = \{1, \dots, n\}$. We assume that each image is illuminated by exactly one point light source and that global and ambient illumination effects are negligible. Moreover, we assume the images to be undistorted, de-vignetted, and the black frame to be subtracted.

We represent the scene as a set of 2D parameter maps defined at several keyframes of the RGB-D sequence. More specifically, at each keyframe, we store the geometry in terms of a depth and normal map and the materials as BRDF parameter maps. Additionally, each



(a) Keyframes (b) Neighbor Views

Figure 5.3: Keyframe and Neighbor View Selection. *(a) Visualized are all keyframes K (red, green) for the object 'Sheep' - they represent the scene as a set of 2.5D maps for efficient optimization. (b) Our method optimizes the parameter set of one keyframe $k \in K$ (green) guided by photometric and geometric constraints from neighboring observation views N_k (blue cameras) and consistency constraints from neighboring keyframes $\bar{N}_k$ (orange lines).*

keyframe is linked to a set of camera poses of its respective neighbor views. In the following, we first describe the process of keyframe and neighbor view selection in Sec. 5.2.1, followed by the representations for poses, geometry, and materials in Sec. 5.2.2.

5.2.1 Keyframe and Neighbor Selection

To represent the scene, we define a set of keyframes $K \subseteq N$ that capture the scene tightly. For each keyframe $k \in K$, we define two sets of neighboring views: The first set is the set of neighboring observation views N_k, which provide photometric and geometric constraints for the local 2.5D multi-view optimizations over the parameter set of keyframe k. Second, we define a set of neighboring keyframes $\bar{N}_k$ from which we project the parameter maps into keyframe k as a soft constraint during optimization to enforce consistency of the local 2.5D reconstructions. These pairwise consistency constraints propagate globally during optimization since all keyframes are connected via the overall optimization graph. As evidenced by our experiments, this term is crucial for obtaining a consistent result when fusing all 2.5D representations into a global 3D representation of the scene. All sets of views are visualized in Fig. 5.3 for the capture of the object 'Sheep'.

Keyframe Selection: To select a set of diverse keyframes $K \subseteq N$, we iteratively compute the pairwise 3D Euclidean distances between the camera centers of all views and remove the view with the minimum distance to its nearest neighbor until the desired number of keyframes has been reached. Generally, the number of keyframes is a tradeoff between accuracy and time, and it grows with the scale of the scene. However, increasing the number of keyframes is unproblematic for our method since most computations run per keyframe in parallel, with fixed memory requirements independent of the scene size. We ablate the number of keyframes in Sec. 5.6.2.

Neighboring Observation Views N_k: The set of keyframe observations contains too few samples to optimize geometry, pose, and spatially-varying material parameters. Therefore, we define m neighboring observation views $N_k \subset N$ with $m = |N_k|$ per keyframe $k \in K$ and minimize the photo-consistency error between these and the predictions of our model. To select the neighbor observation views, we only consider views within a 40° cone around keyframe k with respect to (wrt.) the object center. For larger scenes, we additionally remove views with a view direction that deviates more than 45° from the keyframes' view direction. We then choose m views that cover the cone around keyframe k as uniformly as possible by removing the views closest to their neighbors.

Neighboring Keyframes $\bar{N}_k$: For each keyframe $k \in K$, we define a set of neighbor keyframes $\bar{N}_k \subseteq K \setminus \{k\}$ of size $\bar{m} = |\bar{N}_k|$. During optimization, we regularize the parameter maps of keyframe k against those of all neighbor keyframes $i \in \bar{N}_k$ projected into k. This enforces consistent parameter estimates across keyframes. To ensure that all neighbors $i \in \bar{N}_k$ share scene content with keyframe k, we sample them randomly from all keyframes that fulfill two conditions: For keyframes i and k, 1) define the *middle point* as the median of all initial geometry points for objects and the first intersection point of the principal ray of camera k with the initial geometry for scenes. Then the two lines connecting each view's camera position with the *middle point* should form an angle of $\leq 60^\circ$. And 2) for scenes, both cameras' view directions form an angle of $\leq 45^\circ$. Note that we sample up to $\bar{m}$ neighboring keyframes per keyframe, depending on the availability of valid neighbors.

5.2.2 Keyframe Representations

In this section, we formally describe the keyframe-based parameterization of our model in terms of poses, geometry, and materials. For each keyframe $k \in K$, we define its pixels P_k as the set of all pixels of view k with a non-zero initial depth value. As we bound the depth to be non-negative, this implies $z_p^k > 0$.

Camera Representation

We use the same perspective pinhole camera model and assumptions on calibrated intrinsic camera parameters as in Sec. 4.2.1. Regarding notation, we denote the projective mapping for observation $i \in N_k$ and keyframe $k \in K$ as $\boldsymbol{\pi}_i^k : \mathbb{R}^3 \rightarrow \mathbb{R}^2$. Further, we represent the extrinsic component (camera pose) of this mapping in world coordinates by a unit quaternion $\mathbf{q}_i^k \in SO(3)$ and a translation vector $\mathbf{t}_i^k \in \mathbb{R}^3$. Note that we use a redundant representation (i.e., the camera pose of an observation neighboring multiple keyframes is represented once per keyframe) to enable memory-efficient optimization, one keyframe at a time, while enforcing consistency via additional soft constraints.

Geometry Representation

We use the same geometry representation as in Sec. 4.2.2 – only here, depth and normal maps are defined with respect to a keyframe k. For clarity of notation, we repeat some definitions in the following.

Depth Map: For each pixel $p \in P_k$ of keyframe k at 2D location $\left(u_p^k, v_p^k\right)^T$ and associated depth z_p^k, the 3D point location $\mathbf{x}_p^k$ is given by

$$\mathbf{x}_p^k = \left(\pi_k^k\right)^{-1}\left(u_p^k, v_p^k, z_p^k\right) \tag{5.1}$$

where $\left(\pi_k^k\right)^{-1}$ denotes the inverse projection, which takes a pixel coordinate and depth value and returns the 3D point in world coordinates.

Normal Map: We represent normals as 3D vectors $\{\mathbf{n}_p^k\}_{p\in P_k}$ and, as before, only estimate an angular change wrt. the normal of the previous iteration during optimization.

Material Representation

As in Sec. 4.2.3, we model reflectance properties with svBRDFs and estimate the parameters for all pixels/points $p \in P_k$, which now depend on keyframe k. Compared to the previous method, we implemented some changes detailed in the following.

svBRDF: As in Sec. 4.2.3, we use a modified version of the Cook-Torrance model [CT82] with parameters diffuse albedo $\mathbf{d}_p^k \in \mathbb{R}^3$, specular albedo $s_p^k \in \mathbb{R}$, surface roughness $r_p^k \in \mathbb{R}$, and ignoring the Fresnel effect, as given by Eq. (4.2):

$$\rho_p^k(\mathbf{n}_p^k, \omega_{\text{in}}(\mathbf{x}_p^k), \omega_{\text{out}}(\mathbf{x}_p^k)) = \mathbf{d}_p^k + s_p^k \frac{D(r_p^k)\, G(\mathbf{n}_p^k, \omega_{\text{in}}(\mathbf{x}_p^k), \omega_{\text{out}}(\mathbf{x}_p^k), r_p^k)}{4(\mathbf{n}_p^k \cdot \omega_{\text{in}}(\mathbf{x}_p^k))(\mathbf{n}_p^k \cdot \omega_{\text{out}}(\mathbf{x}_p^k))} \tag{5.2}$$

In contrast to Sec. 4.2.3, we now utilize Disney's GTR model [Bur12] for the microfacet slope distribution $D(\cdot)$, Eq. (2.21) with $\gamma = 2$, and Filament's Smith's function [GA18] for the geometric attenuation factor $G(\cdot)$, Eq. (2.24). This enables rendering our reconstruction results using Filament renderer, [GA18] and [Ham17].

Further, while the previous method used specular base materials to constrain the estimation from sparse reflectance samples, this is not necessary anymore. By promoting consistency between multiple keyframe optimizations, we leverage reflectance information from more observations. This enables modeling the specular BRDF parameters per pixel, yielding a richer and more flexible material representation.

5.3 Optimization Objective

Similar to Sec. 4.3, we jointly optimize geometry, materials, and pose parameters for each keyframe $k \in K$ by minimizing the photometric error between rendered predictions and neighbor view observations while employing multiple additional loss functions for regularization.

For a single keyframe k and its pixels/points $p \in P_k$ and neighbor observation views $i \in N_k$, we wish to estimate the depth z_p^k, geometric surface normals $\mathbf{n}_p^k$, svBRDF parameters $\mathbf{d}_p^k, r_p^k, s_p^k$, as well as the camera poses π_i^k. Denoting the parameter set as

$$\mathcal{X} = \{\{z_p^k, \mathbf{n}_p^k, \mathbf{d}_p^k, r_p^k, s_p^k\}_{p \in P_k}, \{\pi_i^k\}_{i \in N_k}\}_{k \in K}$$

we define our objective function as follows

$$\mathcal{X}^* = \underset{\mathcal{X}}{\operatorname{argmin}} \; \psi_{\mathcal{P}}(\mathcal{X}) + w_{\mathcal{D}}\psi_{\mathcal{D}}(\mathcal{X}) + w_{\mathcal{C}}\psi_{\mathcal{C}}(\mathcal{X}) + \underbrace{w_{\mathcal{GC}}\psi_{\mathcal{GC}}(\mathcal{X}) + w_{\mathcal{N}}\psi_{\mathcal{N}}(\mathcal{X})}_{=\psi_{\mathcal{G}}(\mathcal{X})} + w_{\mathcal{M}}\psi_{\mathcal{M}}(\mathcal{X}) \tag{5.3}$$

For clarity, we often omit the dependency on $\mathcal{X}$ and the relative weights between the individual terms:

$$\mathcal{X}^* = \underset{\mathcal{X}}{\operatorname{argmin}} \; \psi_{\mathcal{P}} + \psi_{\mathcal{D}} + \psi_{\mathcal{C}} + \psi_{\mathcal{G}} + \psi_{\mathcal{M}} \tag{5.4}$$

The individual terms encourage photo-consistency $\psi_{\mathcal{P}}$, depth-consistency $\psi_{\mathcal{D}}$, and multi-view consistency $\psi_{\mathcal{C}}$, impose regularization on the geometry $\psi_{\mathcal{G}}$ by encouraging geometric consistency $\psi_{\mathcal{GC}}$ and normal smoothness $\psi_{\mathcal{N}}$, and enforce material smoothness $\psi_{\mathcal{M}}$.

Optimization Weights: The weights are the same for all scenes. Empirically, we found the following values balance the optimization parameters best: For the first iterations $(0 - T/2)$, we set a relatively high smoothness loss on the normals while keeping the specular parameter smoothness constraint lower. This helps boot-strapping the specular material parameters, which are initialized randomly. The loss weights are as follows: $w_{\mathcal{D}} = 5e6$, $w_{\mathcal{C}} = 1e5$, $w_{\mathcal{GC}} = 1e6$, $w_{\mathcal{N}} = 5e4$, and $w_{\mathcal{M}} = 3e4$.

For the second half of the optimization (iterations $T/2 - T$), we reduce the smoothness constraint on the normals to allow for fine details while increasing the material smoothness regularizer to facilitate the propagation of reflectance information across pixels. The loss weights are $w_{\mathcal{D}} = 5e6$, $w_{\mathcal{C}} = 1e5$, $w_{\mathcal{GC}} = 1e6$, $w_{\mathcal{N}} = 1e3$, and $w_{\mathcal{M}} = 3e5$ for specular albedo, and $w_{\mathcal{M}} = 1e5$ for specular roughness smoothness.

Changes to the Objective in Sec. 4.3: In comparison to the optimization objective of the 2.5D method presented in Chap. 4, the greatest change is the new multi-view consistency loss term. Additionally, we make several improvements to the optimization objective, which we empirically found useful. In detail, we refine the following terms:

1. We downweight observations at grazing angles, as these pixels intrinsically contain a higher information uncertainty.

2. We regularize the depth maps against all neighboring depth maps instead of a single

one to better ensure proximity to the measurements,

3. We include an edge-aware weighting term in the normal smoothness loss to facilitate the reconstruction of small details, and

4. We change the material model to a more flexible and practical solution for larger scenes that does not require material clustering and model selection.

We elaborate on these improvements in the following Sec. 5.3.1 to Sec. 5.3.5.

5.3.1 Photo and Depth-Consistency

We introduce the photo and depth-consistency terms in the following. For better readability, we denote $\mathcal{I}_i(\pi_i^k(\mathbf{x}_p^k))$ as the observation $\mathcal{I}_i$ at the 2D location where 3D point $\mathbf{x}_p^k$ is observed in image i. For fractional image coordinates, we use bilinear interpolation. Similarly, we write $\mathcal{Z}_i(\pi_i^k(\mathbf{x}_p^k))$ for depth measurements.

Photo-Consistency: We use the photo-consistency loss term from Sec. 4.3.1 but down-weight observations at grazing angles: We use a multiplicative weight $w_p^i \propto \mathbf{n}_p^i \mathbf{l}_p^i$ proportional to the angle between the surface normal $\mathbf{n}_p^i$ and light direction $\mathbf{l}_p^i$. Additionally, all parameters now depend on a keyframe k. This alters Eq. (4.5) to

$$\psi_{\mathcal{P}}(\mathcal{X}) = \sum_{i \in N_k} \sum_{p \in P_k} \left\| \varphi_p^{ki} \, w_p^i \left[\mathcal{I}_i(\pi_i^k(\mathbf{x}_p^k)) - \mathcal{R}_i(\mathbf{x}_p^k, \mathbf{n}_p^k, \rho_p^k) \right] \right\|_1 \tag{5.5}$$

It ensures that our models' renderings match the observation $\mathcal{I}_i$ for every neighbor view $i \in N_k$ and all visible and illuminated ($\varphi_p^{ki} = 1$) pixels p. $\mathcal{R}_i$ denotes the rendering equation, as given in Eq. (4.6).

Depth-Consistency: The depth-consistency is similar to the *depth compatibility* term in Sec. 4.3.2. But while Eq. (4.9) regularized the depth estimate $\{z_p\}_p$ against the depth measurements $\mathcal{Z}_1$ of the reference view only, we now constrain the depth estimate $\{z_p^k\}_p$ against the depth measurements $\mathcal{Z}_i$ of all neighboring views $i \in N_k$:

$$\psi_{\mathcal{D}}(\mathcal{X}) = \sum_{i \in N_k} \sum_{p \in P_k} \varphi_p^{ki} \left\| z_p^i - \mathcal{Z}_i(\pi_i^k(\mathbf{x}_p^k)) \right\|_2^2 \tag{5.6}$$

Here, z_p^i denotes the depth of the 3D point $\mathbf{x}_p^k$ of keyframe k when projected to the neighbor view i via $\pi_i^k(\mathbf{x}_p^k)$. As before, φ_p^{ki} ensures that surface point $\mathbf{x}_p^k$ is visible in image i.

5.3.2 Multi-View Consistency

Since our representation is composed of multiple 2.5D views, we must ensure consistency between them. Towards this goal, we augment our objective with a multi-view consistency term which encourages the current parameter estimates $\{z_p^k, \mathbf{d}_p^k, r_p^k, s_p^k\}_{p \in P_k}$, $\{\mathbf{t}_j^k, \mathbf{q}_j^k\}_{j \in N_k}$ of keyframe k to agree with those of neighboring keyframes $\{\bar{z}_p^k, \bar{\mathbf{d}}_p^k, \bar{r}_p^k, \bar{s}_p^k\}_{p \in P_k}$, $\{\bar{\mathbf{t}}^i, \bar{\mathbf{q}}^i\}_{i \in \bar{N}_k}$

projected into the current keyframe:

$$\psi_{\mathcal{C}}(\mathcal{X}) = \frac{1}{|P_k|} \sum_{p \in P_k} \left\| \mathbf{x}_p^k - \bar{\mathbf{x}}_p^k \right\|_2 + \left\| \mathbf{d}_p^k - \bar{\mathbf{d}}_p^k \right\|_1 + \left| r_p^k - \bar{r}_p^k \right| + \left| s_p^k - \bar{s}_p^k \right| + \frac{1}{n} \sum_{i \in \bar{N}_k} \sum_{j \in N_i \cap N_k} \left\| \mathbf{t}_j^k - \mathbf{t}_j^i \right\|_1 + \left\| \left(\left(\mathbf{q}_j^i \right)^{-1} \otimes \mathbf{q}_j^k \right)_v \right\|_1 \tag{5.7}$$

Hereby, the projected neighbor parameters $\{\bar{z}_p^k, \bar{\mathbf{d}}_p^k, \bar{r}_p^k, \bar{s}_p^k\}_{p \in P_k}$ for each surface point $\mathbf{x}_p^k$ are computed as follows

$$(\bar{\cdot})_p^k = \frac{1}{\sum_i \varphi_p^{ki} \, w_p^i} \sum_{i \in \bar{N}_k} \varphi_p^{ki} \, w_p^i \, \text{interp}\left(\{(\cdot)_q^i\}_q, \pi_i(\mathbf{x}_p^k) \right) \tag{5.8}$$

where $\bar{N}_k$ is the set of neighboring keyframes, φ_p^{ki} denotes visibility as defined above, and $w_p^i = \mathbf{n}_p^i \mathbf{l}_p^i$ downweights estimates at grazing angles. The mapping $\text{interp} : \mathbb{R}^{|P|} \times \mathbb{R}^2 \to \mathbb{R}$ takes a neighboring parameter map $\{(\cdot)_q^i\}_{q \in P_i}$ and a projected 2D pixel location $\pi_i(\mathbf{x}_p^k)$, and outputs the bilinearly interpolated parameter value. The poses of all neighboring keyframes $\{\bar{\mathbf{t}}^i, \bar{\mathbf{q}}^i\}_{i \in \bar{N}_k}$ are defined as

$$\bar{\mathbf{t}}^i = \left\{ \mathbf{t}_j^i \mid j \in N_i \cap N_k \right\} \quad \text{and} \quad \bar{\mathbf{q}}^i = \left\{ \mathbf{q}_j^i \mid j \in N_i \cap N_k \right\} \tag{5.9}$$

The $\otimes$ operator in the last term of (5.7) denotes the Hamiltonian product for quaternions and calculates the composed rotation. Furthermore, $(\cdot)_v$ denotes the vector part of the quaternion, which equals zero for the identity. Note that we only calculate the pose loss term for cameras which are part of the observations for both the current keyframe k and the neighboring keyframes i, i.e., $j \in N_i \cap N_k$.

5.3.3 Geometry Regularization

Our geometry regularizers encourage geometric consistency $\psi_{\mathcal{GC}}$ and normal smoothness $\psi_{\mathcal{N}}$ as follows:

$$\psi_{\mathcal{G}} = \psi_{\mathcal{GC}} + \psi_{\mathcal{N}} \tag{5.10}$$

As before, we omitted the dependency on $\mathcal{X}$ and the relative weights of the individual terms for clarity.

Geometric Consistency: We enforce consistency between depth $\{z_p^k\}$ and normals $\{\mathbf{n}_p^k\}$ by maximizing the inner product between the estimated normals $\{\mathbf{n}_p^k\}$ and the cross product of the surface tangents at $\{\mathbf{x}_p^k\}$ just as we did in Sec. 4.3.2. Please refer to Eq. (4.7) and Eq. (4.8) for details.

Normal Smoothness: We further encourage normals $\mathbf{n}_p^k$ and $\mathbf{n}_q^k$ of adjacent pixels $p \sim q$ to be similar:

$$\psi_{\mathcal{N}}(\mathcal{X}) = \sum_{p \sim q} e_{pq}^k \left\| \mathbf{n}_p^k - \mathbf{n}_q^k \right\|_1 \tag{5.11}$$

In contrast to Eq. (4.10), here, we apply an edge-aware weighting term e^k_{pq} based on a Canny filter [Can86]. This reduces the smoothing at pixels close to edges in the albedo map and facilitates detailed geometry reconstruction.

5.3.4 Material Smoothness

To enforce propagation of reflectance parameters across pixels, we constrain the specular albedo $\{s^k_p\}_p$ and roughness $\{r^k_p\}_p$ maps against a bilaterally smoothed version of themselves:

$$\begin{aligned}\psi_{\mathcal{M}}(\mathcal{X}) = \sum_p &\left\| s^k_p - \frac{\sum_q s^k_q w^k_q g^k_{pq}}{\sum_q w^k_q g^k_{pq}} \right\|_1 \\ &+ \left\| r^k_p - \frac{\sum_q r^k_q w^k_q g^k_{pq}}{\sum_q w^k_q g^k_{pq}} \right\|_1 \end{aligned} \tag{5.12}$$

The Gaussian kernel g^k_{pq} is given by Eq. (4.12), with both the 3D location $\mathbf{x}^k$ and diffuse albedo $\mathbf{d}^k$ at pixels p and q as features.

This regularization is very similar to Sec. 4.3.3, both assuming that nearby pixels with similar diffuse behavior also exhibit similar specular behavior. However, as we used specular base materials in the previous 2.5D method, we constrained the per pixel specular BRDF weights α_p to be smooth and sparse, Eq. (4.11). Now, without using specular bases, we directly apply the smoothness regularizer to the specular parameter maps and omit the sparsity constraint, Eq. (5.12).

5.3.5 No Pose Regularization

We do not apply any additional pose regularization as the combination of multiple loss terms guides our pose optimization: Foremost, each pose is constrained by the RGB and (rough) depth observation of its respective view and all neighboring observations views via the photo-consistency and depth-consistency regularizers. Any pose drift of individual views results in projection errors for neighboring observations and is penalized by our loss. Further, the parameter predictions of all neighboring keyframes (including the pose parameters) guide the optimization of each keyframe via the multi-view consistency loss. Given these constraints in our objective function, we do not observe any large drift of any individual pose in our experiments.

A joint drift of all poses along the same vector or angle is conceivable. In this case, the full reconstruction would change its location and orientation in world space. The chance of this happening is low since the multi-view consistency regularizer is updated every $t = 100$ iterations. Thus, poses are regularized against the temporarily fixed neighbor keyframe poses, which hampers global drift. If all poses drifted globally, the reconstruction results would not be degraded – without loss of generality, any camera coordinate system can be considered as the world coordinate system.

Data: Color and depth images $\{\mathcal{I}_i, \mathcal{Z}_i\}_{i \in N}$.
Result: Mesh $\mathcal{M}$ featuring per-vertex normals and BRDF parameters.

Initialize $\forall k \in K$: // Sec. 5.4.3
$\mathcal{X}_0^k = \{z_p^k, \mathbf{n}_p^k, \mathbf{d}_p^k, r_p^k, s_p^k\}_{p \in P_k}$ and $\{\pi_i^k\}_{i \in N_k}$
$\bar{\mathcal{X}}_0^k = None$

$t = 100,\ T = 2000$
$rounds = T/t$

Multi-View Consistent Optimization:
for $r = 1$ **to** *rounds* **do**
  for *keyframe* $k \in K$ **do**
    Optimize $\mathcal{X}_{r-1}^k$ given $\bar{\mathcal{X}}_{r-1}^k$ for t iterations: // Eq. (5.4)
    $\mathcal{X}_r^k = \mathcal{X}_{r-1}^{k,*}$
    Project neighbor parameter maps: // Eq. (5.8)
    $\bar{\mathcal{X}}_r^k = \{\bar{z}_p^k, \bar{\mathbf{d}}_p^k, \bar{r}_p^k, \bar{s}_p^k\}_{p \in P_k}$ and $\{\bar{\pi}_i^k\}_{i \in \bar{N}_k}$
  end
end

Mesh Generation: // Sec. 5.5
  Fuse all final keyframe parameter maps $\{\mathcal{X}_{r=T/t}^k\}_{k \in K}$
  into a mesh $\mathcal{M}$ by volumetric fusion and marching cubes.

Algorithm 1: Pseudocode of the Proposed Algorithm.

5.4 Optimization

Direct optimization of the global objective in Eq. (5.4) does not scale to larger scenes due to the large amount of data (variables and observations) that need to be stored and GPU memory limitations. Instead, we decompose the global reconstruction into multiple keyframe reconstructions and perform decentralized, frame-wise block coordinate descent in parallel on multiple processes. With this distributed optimization strategy, we drastically reduce the memory footprint since we only ever need to store one block in memory at a time per process. To keep locally adjacent blocks consistent, we periodically share the state of optimization variables between neighboring keyframes and regularize differences in the reconstructed models. An overview of the full optimization algorithm is given in Algo. 1.

In the following, we first motivate our decentralized optimization strategy in Sec. 5.4.1 and then elaborate on the block coordinate descent algorithm in Sec. 5.4.2. Subsequently, we provide details about the parameter initialization in Sec. 5.4.3 and our implementation in Sec. 5.4.4.

5.4.1 Decentralized Optimization

For large computational problems, a **distributed optimization strategy** that allows for multiple processes and parallelization is essential. As per [Ass+20], distributed methods can be categorized into centralized and decentralized algorithms, depending on whether the processes read and update a central copy of the optimization variables or work on independent local copies. Centralized algorithms require consistent transaction management, such as semaphores, cause additional computational costs for centralization and distribution, and exhibit less stable optimization behavior due to potentially contradicting updates to the central optimization variables by different processes. Therefore, we implement a decentralized algorithm that facilitates accurate local reconstructions. As this might result in multiple different estimates per variable, we introduce a soft regularizer to establish synchronization between processes and encourage consistency across spatially nearby regions. We reduce the communication overhead by employing a strategy similar to [Wan+20] and letting each process independently perform a set of base optimization steps in between synchronization. In the following, we call this set of optimization steps performed for all processes a '*round*'.

5.4.2 Block Coordinate Descent Optimization

For optimization, we represent the target scene as a set of 2.5D parameter maps induced by the keyframes. The keyframes can be viewed as blocks of variables over which we iterate, as described in the following:

We start the first round by optimizing over every block/keyframe for t iterations independently using gradient descent. Given these intermediate optimization states $\{\mathcal{X}_1^k\}_k$ of all keyframes in K, the current parameter map estimates are projected into neighboring keyframes. We refer to them as $\{\bar{\mathcal{X}}_1^k\}_k$. They are used in the subsequent round when the parameters of all blocks/keyframes are re-optimized for t iterations with the additional multi-view consistency regularizer. We iterate these rounds of optimizing for $\{\mathcal{X}_r^k\}_k$ and calculating $\{\bar{\mathcal{X}}_r^k\}_k$, for a total of T/t rounds and T iterations. During the first half of optimization, we use a higher geometric smoothing regularizer to help bootstrap the parameter maps. Thereafter, the smoothness regularization is reduced to allow for carving out fine geometric details and modeling sharp specular highlights during the remaining iterations.

5.4.3 Initialization

We initialize the **poses** using the SfM pipeline COLMAP [SF16; Sch+16]. For the **depth** initialization, we pre-integrate the relatively coarse data of an active stereo setup into a fused 3D geometry using volumetric fusion [Zen+17] and render an initial depth map per keyframe $k \in K$, see Fig. 5.9 for an example of the initial geometry.

Initial **diffuse albedo** and **normal** maps can be computed in closed form assuming a Lambertian scene, as described in Sec. 4.4.1. Both **specular albedo** and **roughness** parameter maps are initialized by sampling randomly and uniformly from the intervals $[0.05, 0.25]$ and $[0.1, 0.9]$, respectively.

5.4.4 Implementation Details

We have implemented the rendering function $\mathcal{R}_i$ using PyTorch [Pas+17], exploiting PyTorch's GPU acceleration and auto-differentiation capabilities. For our experiments, we use $|K| = 24$ keyframes, $m = 20$ neighboring observation views, and $\bar{m} = 10$ neighboring keyframes. We project parameter maps into neighboring keyframes every $t = 100$ iterations and optimize for $T = 2000$ iterations in total.

Optimizer and Learning Rates: We use the Adam optimizer as provided in Pytorch with weights $\beta = [0.9, 0.9]$ and $\varepsilon = 1e-3$. The learning rates are defined per parameters: Let $k \in K$ denote the keyframe, $i \in N_k$ the neighbor observation views, and $p \in P_k$ its pixels/points. We set the learning rate for depth z_p^k and camera poses π_i^k to $1e-4$, for the surface normals $\mathbf{n}_p^k$ to $\frac{\pi}{360}$, and for all svBRDF parameters $\mathbf{d}_p^k, r_p^k, s_p^k$ to $1e-2$. For scheduling, we multiply the learning rates by 0.1 after $\frac{1}{3}$ and $\frac{2}{3}$ of 1000 iterations.

5.5 Mesh Generation

In order to obtain a full 3D reconstruction of geometry and materials, we use a memory-efficient, voxel hashing-based implementation of volumetric fusion [CL96] as seen in [Nie+13]. Since predictions are most accurate for pixels observed frontally, we weight each pixel contribution by the cosine between the surface normal and viewing direction. We extract the final mesh from the TSDF via marching cubes [LC87]. As we impose consistency across keyframes during optimization, the fused parameter maps are consistent without the need for extra post-processing/alignment steps. Finally, the resulting mesh allows for extracting a texture map for each svBRDF parameter using Blender and exporting the mesh into the OBJ file format. In all our experiments, we use a voxel size of 0.5 - 2mm.

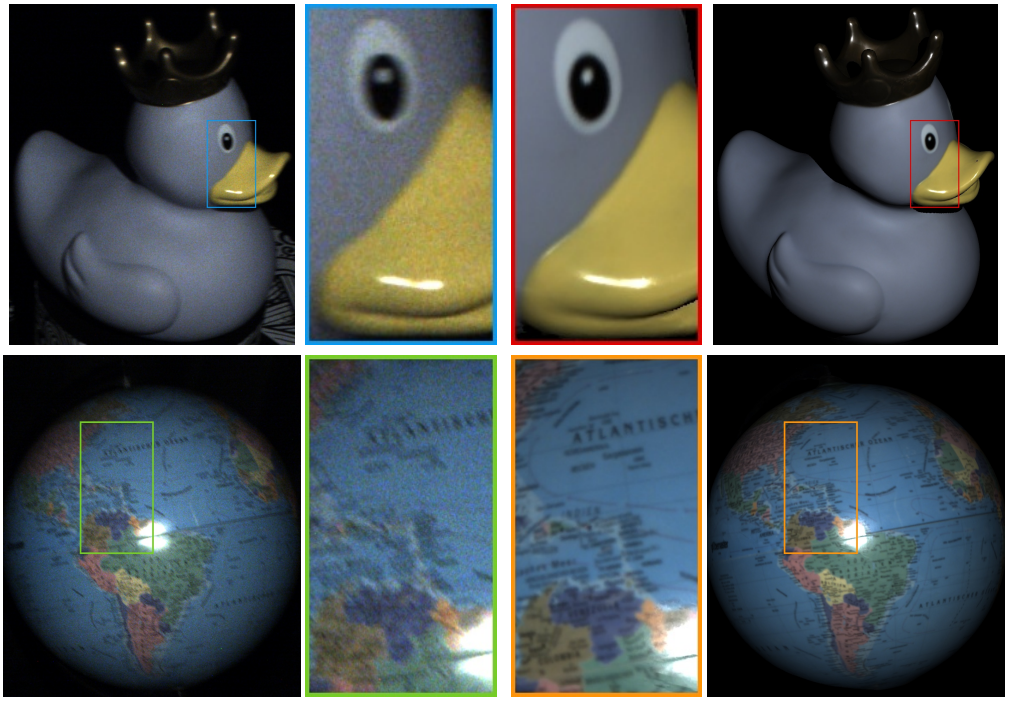


Figure 5.4: Super-Resolution and Denoising (3D). *With a handheld capture system, the measured observations exhibit image noise and motion blur (left), which our model is able to remove. The resulting reconstructions appear denoised and sharpened (right).*

5.6 Experimental Evaluation

The method presented in this chapter is an extension of the 2.5D reconstruction algorithm presented in the previous chapter to full 3D models. We refer to this earlier method from Chap. 4 as '*our 2.5D method*' or '*Previous*' and call the extension of the current Chap. 5 '*our 3D method*' or '*Proposed*'.

In this section, we demonstrate that the simple fusion of multiple 2.5D reconstructions obtained using the previous method is insufficient to obtain accurate 3D models. In particular, the geometry from different keyframes does not align well enough, and material predictions from different viewpoints often differ noticeably due to the ambiguities present in this inverse problem. We show that a multi-view consistent optimization scheme enables consistent and accurate reconstructions of 3D models. We further demonstrate that it allows for modeling larger scenes beyond single objects.

In the following, we present the results of our 3D method and provide an evaluation of captures of real objects and scenes from our custom-built handheld sensor rig. First, we describe the data capture procedure and then provide details on our evaluation protocol in Sec. 5.6.1. Afterward, we conduct an ablation study of our method's components in Sec. 5.6.2 and provide qualitative and quantitative comparisons to related approaches in

Sec. 5.6.3. We then present our method's results on synthetic data in Sec. 5.6.4 and on our captured real dataset in Sec. 5.6.5. Sec. 5.6.6 concludes with a discussion of our method's limitations and an outlook. Note that unlike in Chap. 4, all of the following results show fused 3D models unless explicitly stated otherwise.

5.6.1 Setup

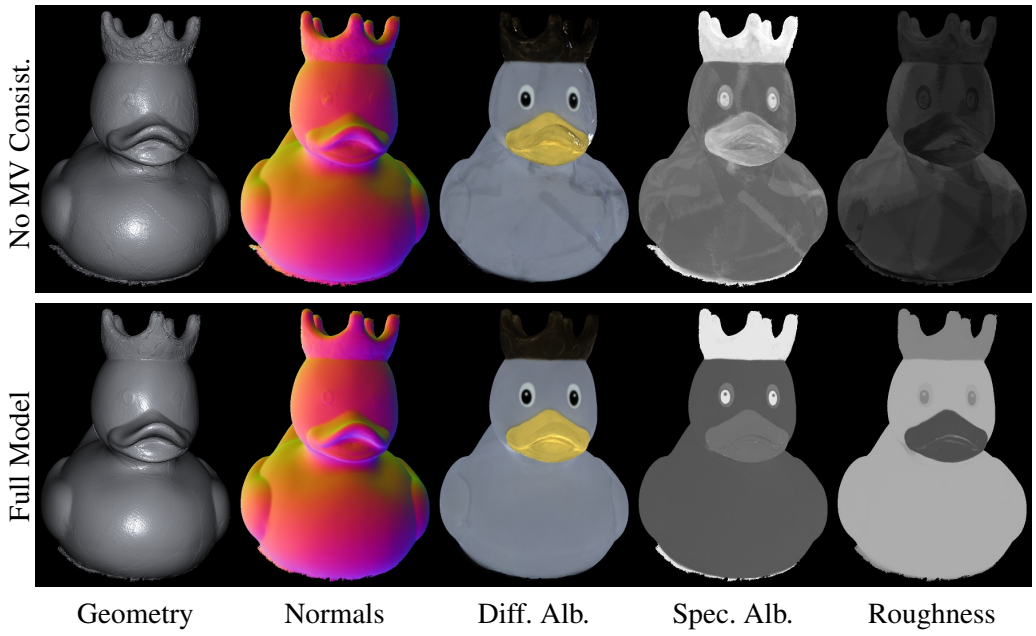
We use the custom-built handheld sensor rig presented in Sec. 4.5.1 to capture data. Examples of captured raw observations are shown in Fig. 5.4 (left). Note that due to the handheld setup, we need to accept a certain amount of image noise to trade off motion blur. However, similar to Fig. 4.4, we show in Fig. 5.4 (right) that our 3D method is able to predict denoised and sharpened reconstructions. If the scene is not sufficiently textured, we additionally add texture patterns to the scene to ensure enough feature points and obtain more reliable initial pose estimates. We use full image resolution (4K) for all objects and half image resolution (2K) for scenes due to GPU memory limitations. We captured between 800 - 1400 images for all our objects and scenes.

Evaluation Protocol: For quantitative evaluation, we render the final model in 10 held-out test views and compute the photometric loss with respect to test observations.

Our model can deviate from the coarse initial COLMAP poses during optimization, so we first align the test view poses to the predicted model. Next, we compute the Root Mean Square Error over all pixels with a non-zero color prediction for each test view and report the mean loss. Since for real capture data, there exist no ground truth object masks, we define valid image pixels as pixels with a non-zero prediction (prediction mask). For validation, we draw the observation masks by hand for multiple objects and find that the RMS errors for both evaluation masks differ by $< 5\%$. Therefore, for simplicity, we use the prediction mask in the following for all experiments.

	No MV Consistency	Full Model
Photometric Test Error	14.65	13.295

(a) *RMSE on held-out test views, averaged over three objects.*



(b) *Qualitative comparison of estimated parameter maps.*

Figure 5.5: The Multi-View Consistency Loss (3D) *facilitates consistent parameter predictions across keyframes resulting in more accurate reconstructions (a). In (b), we show parameter maps on the 3D fused mesh. Without the multi-view consistency regularizer (top), geometric artifacts are visible, and the BRDF parameters show patch-like structures as well as baked-in shading on very glossy object parts (e.g., beak and crown). In contrast, with our loss (bottom), the mesh is clean and materials are homogeneous per object part.*

5.6.2 Ablation Study

In this section, we ablate the important parts of our model, qualitatively and quantitatively.

Multi-View Consistency Loss: Multiple local reconstructions of our method share the coarse but consistent captured depth maps but observe different samples of the reflectance function since the captured images contain only sparse measurements thereof. That implies that while the 2.5D keyframe optimizations lead to consistent geometry with respect to immediate neighbors, they may lead to inconsistencies wrt. keyframes that are not optimized. And these inconsistencies cannot be resolved reliably using volumetric fusion, which merely averages multiple geometries. Thus, there is no guarantee for consistent reconstruction results without explicitly enforcing consistency between keyframes.

	Fixed Poses	Full Model
Photometric Test Error	18.767	17.8035

Table 5.1: Pose Optimization (3D). *Similar to our conclusions in 2.5D (Fig. 4.6), pose optimization also improves the fused 3D results of our full model.*

Fig. 5.5 demonstrates the effectiveness of the multi-view consistency loss. It encourages both regularization and propagation between keyframes. Synchronization of parameter estimates during optimization enables our method to find an equilibrium of the variables which is consistent with not only its neighboring observation views but also the neighboring observation views of nearby keyframes. These connections between neighboring keyframes form a connected graph over all keyframes. Therefore, consistency between any two keyframes can be penalized during optimization, which enforces global consistency. Without the multi-view consistency loss term, the fused model is not aligned well enough to form one coherent surface or consistent BRDF parameter maps. In contrast, we observe that our distributed multi-view consistent optimization leads to globally consistent results without blending artifacts as illustrated in Fig. 5.5b. We note that, in particular, specular properties (specular albedo and roughness) are robustly reconstructed despite the sparsely sampled reflectance function and the high sensitivity of specular highlights to angular configurations. This leads to lower reconstruction errors as evidenced in Tab. 5.5a.

Pose Optimization: Misaligned camera poses yield wrong correspondences between view pairs and can cause various reconstruction artifacts such as ghosting, blur, and texture/ geometry bleeding between front and back surfaces. With respect to material reconstruction, wrong pose estimates lead to errors in the prediction of angular relations between the surface normals, view, and light directions. This causes estimated specular highlights not to align with the mirror reflection direction and hence leads to highlights not being recovered – bake-in effects of specular appearance into predicted texture and normal maps are the result. Such pose alignment problems are particularly crucial when working with a moving handheld scanner. Therefore, we optimize the camera poses jointly with the other parameters leading to more consistent results and lower reconstruction errors. We show these findings quantitatively for our fused 3D method in Tab. 5.1, confirming the results of our 2.5D method in Fig. 4.6.

Geometry and Material Loss Weights: Balancing normal and material smoothness loss terms is crucial for plausible and accurate results. We conduct a series of experiments both qualitatively and quantitatively to ablate the normal and material smoothness loss weight terms, see Fig. 5.6, Fig. 5.7, and Tab. 5.2.

Results show that too little guidance by the material smoothness term prevents material information from propagating across pixels. But this propagation is crucial since a mobile sensor only provides sparse BRDF measurements that do not suffice to constrain the material estimation. Incoherent material maps and baked-in specularities are the results, i.e., see Fig. 5.6 (orange boxes). Contrary, with too much material smoothing, any variation in the material parameters is penalized, predictions over all pixels are drawn together, and

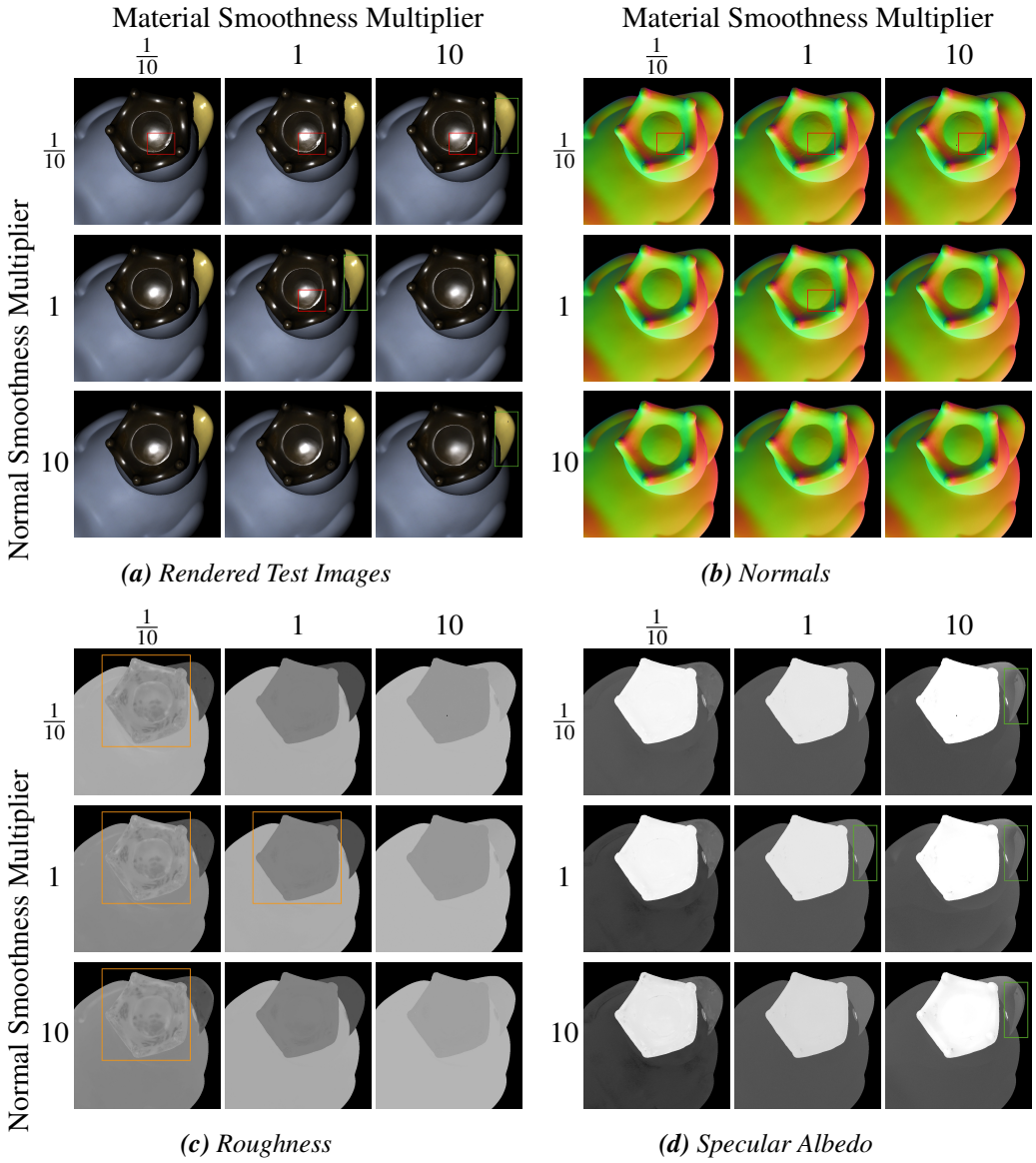


Figure 5.6: Qualitative Ablation of Loss Weights (3D) *for the normal and material smoothness regularizers. Results are shown for the object 'Duck'. We ran nine full reconstructions, each with modified loss weights for normal smoothness (differences highlighted in red), material smoothness (differences highlighted in orange and green), or both. With multiplication factors* 10 *and* $\frac{1}{10}$*, we show the full* 3×3 *grids (normal smoothness on the vertical and material smoothness on the horizontal axis) for the predicted normal and material parameter maps as well as a rendered test image. Please find a detailed discussion of the results in the text, Sec. 5.6.2. We recommend the reader to zoom in for details.*

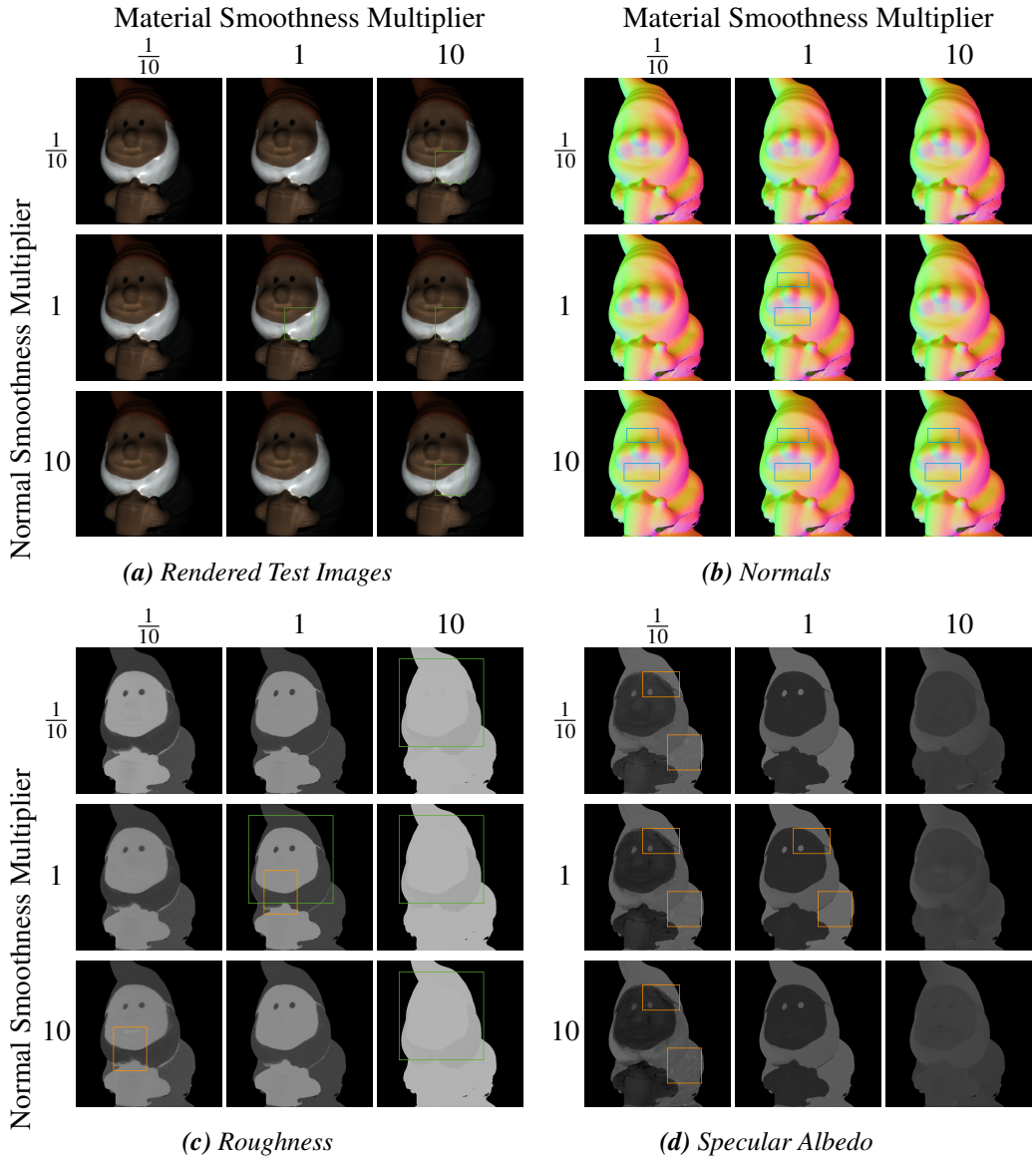


Figure 5.7: Qualitative Ablation of Loss Weights (3D) *for the normal and material smoothness regularizers. Results are shown for the object 'Gnome'. We ran nine full reconstructions, each with modified loss weights for normal smoothness (differences highlighted in blue), material smoothness (differences highlighted in orange and green), or both. With multiplication factors* 10 *and* $\frac{1}{10}$*, we show the full* 3×3 *grids (normal smoothness on the vertical and material smoothness on the horizontal axis) for the predicted normal and material parameter maps as well as a rendered test image. Please find a detailed discussion of the results in the text, Sec. 5.6.2. We recommend the reader to zoom in for details.*

		Material Smoothn. Mult.		
		$\frac{1}{10}$	1	10
N. S. M.	$\frac{1}{10}$	5.91	6.117	5.995
	1	**5.901**	5.949	6.11
	10	6.163	6.152	6.299

(a) *Multiplication Factor* 10

		Material Smoothn. Mult.		
		$\frac{1}{100}$	1	100
N. S. M.	$\frac{1}{100}$	6.679	**6.292**	8.966
	1	6.672	6.315	8.732
	100	8.366	7.696	9.331

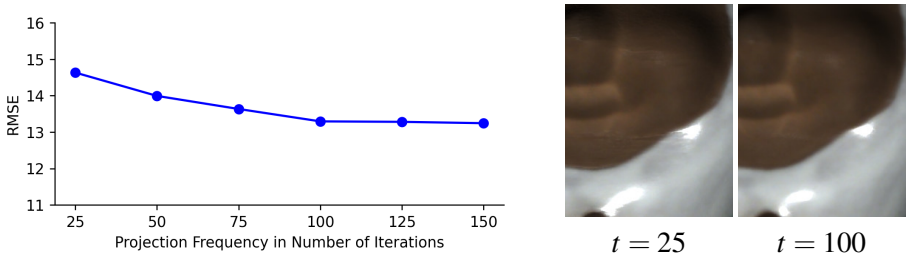
(b) *Multiplication Factor* 100

Table 5.2: Quantitative Ablation of Loss Weights (3D) *for the normal and material smoothness regularizers. We ran nine full reconstructions for several objects and two multiplication factors, each with modified loss weights for normal smoothness, material smoothness, or both. Hereby, (a) shows results for multiplication factors* 10 *and* $\frac{1}{10}$ *and (b) for multipliers* 100 *and* $\frac{1}{100}$*. For both experiments, we show the full* 3×3 *grid of averaged reconstruction losses with the normal smoothness term on the vertical axis and material smoothness on the horizontal axis. Shown is the* L_1 *loss, evaluated per object on 10 held-out test views and averaged over four objects for (a) ('Duck', 'Gnome', 'Globe', 'Sheep') and two objects for (b) ('Gnome', 'Sheep'). Please see the text in Sec. 5.6.2 for a discussion of the results.*

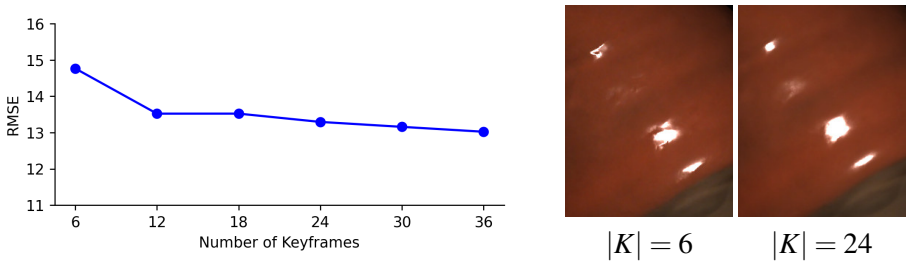
specular highlights disappear from the renderings (Fig. 5.7, green boxes). Furthermore, inconsistencies between parameter estimates of different keyframes increase and overpower the global consistency regularizer, causing additional artifacts as seen in Fig. 5.6 (green boxes). Concerning the normal smoothness loss weights, results show that little smoothing allows for shading effects to appear in the normal maps instead of the material maps. Even small artifacts in the normals cause incorrect specularity predictions in the rendered images, as seen in Fig. 5.6 (red boxes). Conversely, overly smoothed normals blur details, e.g., Fig. 5.7 (blue boxes). This noticeably hampers material estimation, and we observe enhanced errors for, e.g., too little material guidance.

These qualitative findings are confirmed by the quantitative measures in Tab. 5.2. Even though artifacts in normals or specular highlights usually only affect a small portion of the image pixels, we observe favorable results for our configuration of geometry and material smoothness weights. Overall, our experiments show that the loss terms need to be weighed carefully and demonstrate that a good balance for plausible reconstruction results exists.

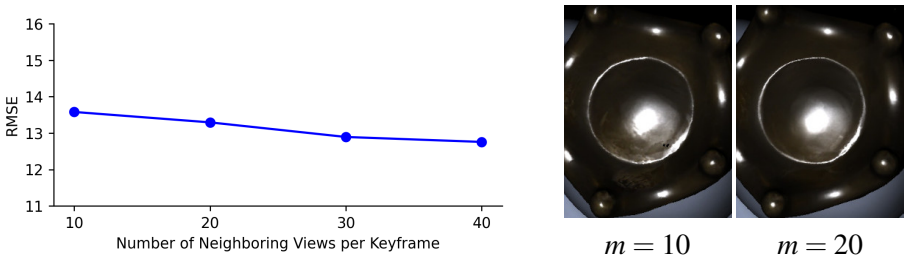
Multi-View Parameter Projection Frequency: In a single optimization round, we optimize the parameters of all keyframes for t iterations before synchronizing with neighboring keyframes. Therefore, t balances local reconstruction quality and global parameter consistency. A low number of t or high synchronization frequency hinders the local optimizations to fit the neighboring observations as shown in Fig. 5.8a (right), whereas a low frequency or no synchronization ($t = T$) prevents consistency among the *current* parameter estimates of neighboring keyframes, as discussed in Fig. 5.5. We found that $t = 100$ leads to accurate parameter estimates and consistent results across keyframes.



***(a)** Multi-View Projection Frequency*



***(b)** Number of Keyframes*



***(c)** Number of Neighboring Observation Views per Keyframe*

Figure 5.8: Multi-View Optimization (3D). *For multiple parameters, we show the average error over three objects wrt. the parameter on the left and example predictions on the right. Here, the left image is a rendered result for the worst choice of parameters, and the right image shows a rendered result for our chosen parameters.*

Number of Keyframes: Fig. 5.8b plots test accuracy against the number of keyframes $|K|$. We observe that generally, more keyframes lead to more accurate reconstructions, and most affected by a small number of keyframes is the quality of the predicted highlights. This offers several insights as it indicates that 1) local keyframe results are globally consistent also for larger numbers of keyframes, 2) details are preserved, and the geometry is not noticeably blurred during mesh fusion and 3) synchronization with neighboring keyframes is essential for correct reflectance estimation (with a reduced number of keyframes, the number of possible neighboring keyframes decreases as well). We use $|K| = 24$ in the following for all single objects as the performance gain becomes very small after that.

Number of Neighboring Observation Views: We aim to estimate the spatially-varying BRDF from only a sparse set of observation samples for each surface point $\mathbf{x}$. As expected, Fig. 5.8c shows that a lower number of neighboring observation views m worsens this

problem. However, this effect is quite small, and the error degrades gracefully. We attribute this to our multi-view consistent optimization scheme, which regularly provides information from neighboring keyframes during optimization. Therefore, given a sufficiently large number of keyframes, our method produces accurate predictions already for $m = 20$.

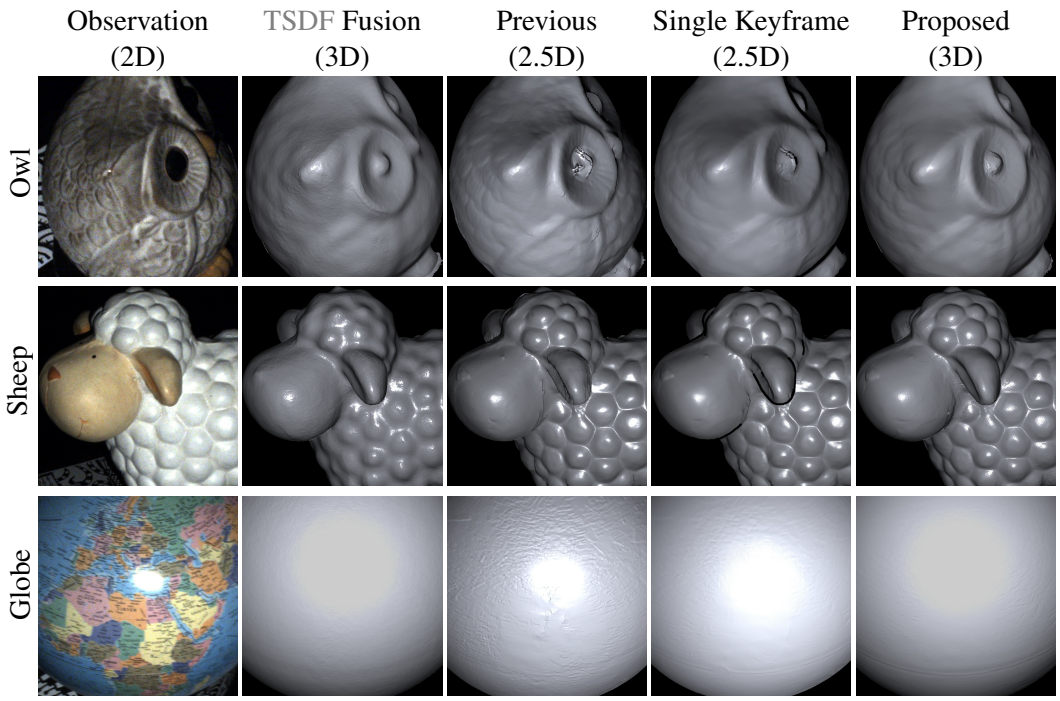


Figure 5.9: Qualitative Geometry Comparison (2.5D vs. 3D). *We show, for each object, the rendered depth map (shaded based on estimated surface normals) for the 3D model after TSDF fusion [Zen+17], our previous 2.5D method, and both a 2.5D keyframe as well as the full 3D model of our proposed 3D method. We observe that the photometric approaches recover more details than naive TSDF fusion of the input geometry. Thanks to our multi-view consistent optimization scheme, the single keyframe result of the proposed 3D method contains fewer textural artifacts in the geometry than our previous 2.5D method (see, e.g., the 'Globe'). Further, we observe that for our model, the resulting global 3D geometry is as detailed as the geometry of the 2.5D keyframe and additionally resolves artifacts present in the local reconstruction, i.e., the eye of the 'Owl'.*

5.6.3 Comparisons to Existing Approaches

We compare this chapter's model with TSDF fusion [Zen+17], our 2.5D method from the previous chapter, and the 3D reconstruction method from Nam et al. [Nam+18]. We qualitatively evaluate our reconstructions regarding geometric details, material modeling, and overall appearance prediction.

Geometry Reconstruction: In Fig. 5.9, we compare the geometry reconstruction capabilities of this chapter's 3D method to the previous chapter's 2.5D method and naive 3D TSDF fusion [Zen+17] of the raw depth maps. We show that both photometric approaches are able to recover fine geometric structures that are not present in the initial reconstruction. Further, in contrast to our 2.5D method, the proposed 3D method recovers the geometry for shiny and dark surfaces like the eyes of the 'Owl'. Such materials are very challenging

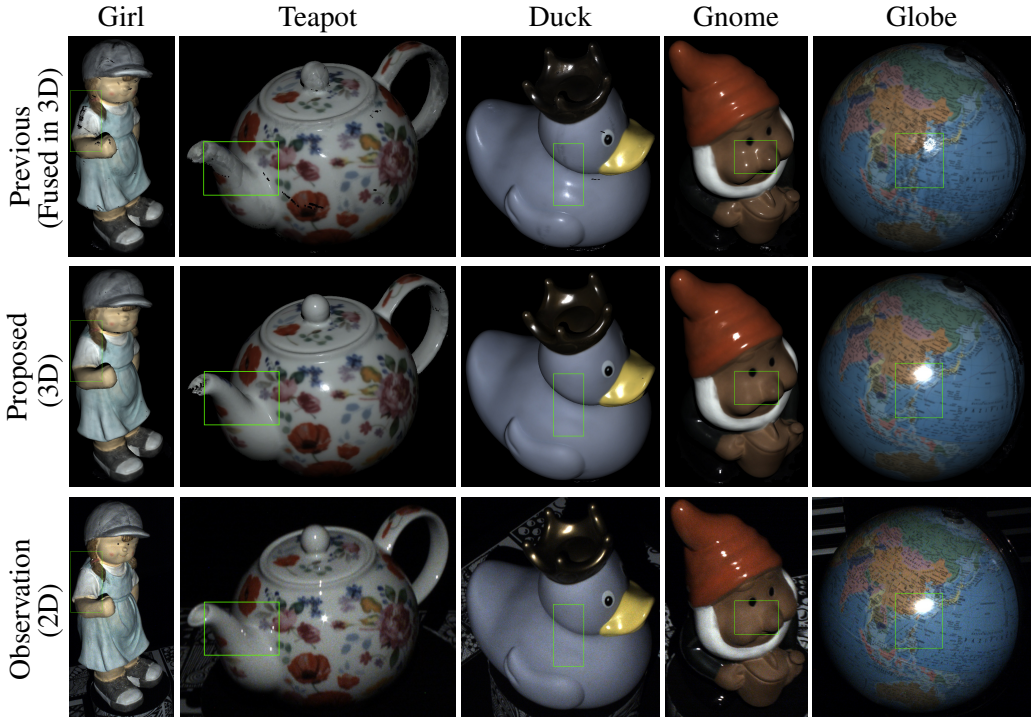


Figure 5.10: Qualitative Comparison to Chap. 4 (3D). *We execute the proposed 3D method and the independent 2.5D reconstructions of our previous method from Chap. 4 for the same set of keyframes. We show both models after volumetric fusion on held-out test views. While our proposed 3D method leads to realistic appearance reconstructions, the predictions of our 2.5D method (independent optimizations) cannot resolve inconsistencies between keyframes and tend to overestimate specular parameters (see highlighted regions of 'Duck', 'Gnome' and 'Girl'). Please see Fig. 5.11 for details and Tab. 5.3 for quantitative results.*

for photometric approaches since the signal-to-noise ratio of the diffuse component is low, and the signal from specular highlights is very sparse. Since our multi-view consistent optimization shares information between keyframes, it is often able to reconstruct such problematic regions.

Comparison to Chap. 4: We present a qualitative and quantitative comparison to our 2.5D method presented in Chap. 4. Since the method jointly reconstructs pose, geometry, and materials for local 2.5D scene representations, we execute it independently on all keyframes in K. We then demonstrate that integration of the results to a fused, global 3D mesh is insufficient: The 2.5D parameter estimates are inconsistent, causing patching artifacts and wrong appearance in the predictions of the integrated 3D models; see Fig. 5.10 and Fig. 5.12. The integration in 3D also reveals ambiguities in the estimation of geometric and photometric parameters. Our case study in Fig. 5.11 shows that the 2.5D method is indeed prone to overestimate specular reflection: Since the appearance is highly sensitive to angular changes in the normals, very small deviations of the normals are sufficient to strongly alter the

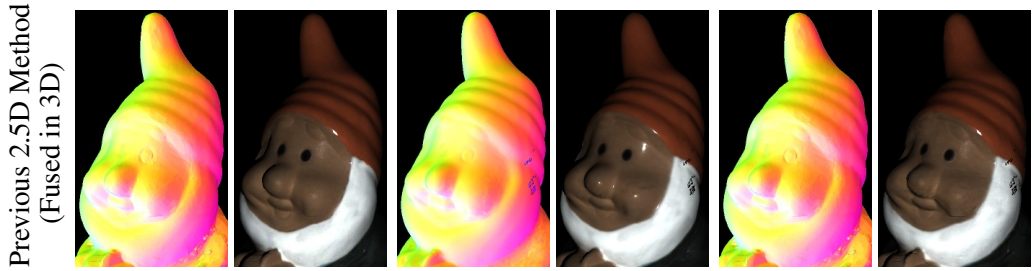


Figure 5.11: Reconstruction Ambiguities for Chap. 4 (2.5D vs. 3D). *Integration of independent 2.5D reconstructions, as presented in the previous Chap. 4, into a fused 3D mesh leads to incorrect appearance predictions, see Fig. 5.10 and Tab. 5.3. This can be attributed to unresolved ambiguities between geometry and materials: Shown are the normal maps and rendered predictions for (left) a single 2.5D reconstruction of our previous method, (middle) the fused 3D mesh after integration of independent 2.5D reconstructions as shown in Fig. 5.10, left column and (right) the fused 3D mesh (as in the middle) with only the normal map loaded from the 2.5D keyframe (from the left). We observe that the appearance of the 3D mesh (middle) shows artifacts. But when using the noisier normal map from the 2.5D keyframe reconstruction, these artifacts are reduced noticeably (right). That indicates that our previous 2.5D method cannot resolve ambiguities in the normal and specular material estimation. It tends to overestimate specular parameters (e.g., on the face of the 'Gnome') and slightly perturbs the normal maps as compensation, resulting in a good appearance.*

	Girl	Teapot	Duck	Gnome	Globe
Previous (Fused in 3D)	15.196	43.755	31.212	64.032	92.765
Proposed (3D)	12.281	18.128	10.366	17.483	10.510

Table 5.3: Quantitative Comparison to Chap. 4 (3D). *In Fig. 5.10, we show qualitative results for the proposed 3D method and for the fused volume of independent 2.5D reconstructions of our previous method presented in Chap. 4, for the same set of keyframes. Here, we present the photometric test error on held-out test views for both these models. Results show that the inconsistencies between keyframes and overestimated specular parameters of the independent optimizations by our 2.5D method, as shown in Fig. 5.10 and Fig. 5.11, lead to significantly higher test errors compared to the proposed 3D method.*

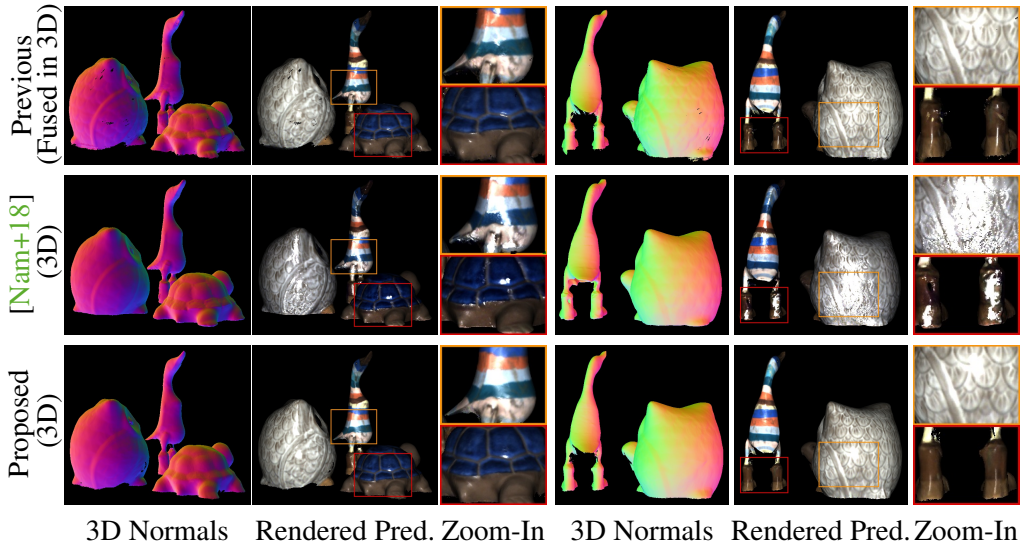
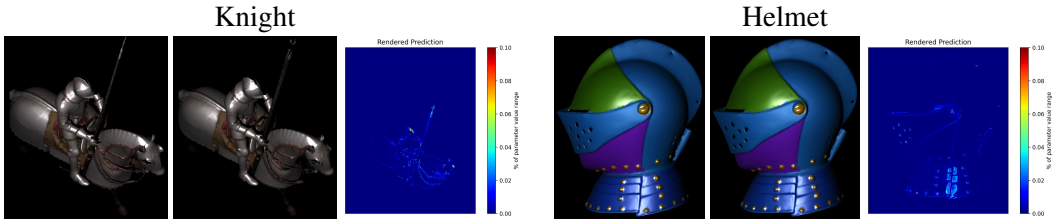


Figure 5.12: Qualitative Comparison to Baselines (3D). *The simple fusion of the 2.5D reconstruction results of our previous method results in artifacts in the geometry and appearance. While the model from Nam et al. [Nam+18] recovers detailed normal maps, the material reconstruction fails to capture the object's appearance, leading to high-frequency artifacts in the predictions. Our 3D model estimates normals that are on par with the results of Nam et al. [Nam+18] and predicts consistent appearance with realistic specular reflections.*

appearance, e.g., by removing specular highlights from the predictions. Therefore, the model is able to '*cheat*' by tilting normals away instead of decreasing the glossiness of the material. In contrast, our 3D method resolves such ambiguities by incorporating information from neighboring keyframes and encouraging multi-view consistency. Specifically, regularization against aggregated neighboring parameter maps prevents a bias in the normals and leads to better material estimates. Tab. 5.3 confirms significantly lower reconstruction errors for our model on held-out test views.

Comparison to Nam et al. [Nam+18]: We compare our method with that of Nam et al. [Nam+18] on a scene with multiple objects. As shown in Fig. 5.12, their method reconstructs 3D appearance components such as normals and diffuse albedo properly but does not recover the specular reflections of the given scene well. Since they use base materials for reflectance modeling, this indicates that the clustering into surface regions with distinct materials fails, potentially due to an erroneous estimate of the number of base materials. This leads to non-smooth predictions even for regions with similar appearance, resulting in uneven specular highlights and high-frequency artifacts. In contrast, our method reconstructs materials and specular highlights well because our pixel-wise material representation does not involve a model selection step.



(a) Initial (b) Fused (c) Error (a) Initial (b) Fused (c) Error

Figure 5.13: Synthetic Ground Truth (3D). *(a) Starting from an initial mesh, we render synthetic observations and parameter maps. Within our reconstruction pipeline, we re-mesh by fusing the reconstructed parameter maps of all keyframes into one consistent 3D mesh. This mesh typically differs from the initial mesh in its vertex count, resulting in an expected difference between renderings of both meshes. (b) To quantify this difference for the synthetic data, we fuse the initial ground truth parameter maps of all keyframes using the same fusion technique we use in our method, hence yielding a lower bound on the attainable reconstruction error. This results in the fused ground truth mesh, which we render into fused observations and parameter maps. (c) Comparing the synthetic observations rendered from both ground truth meshes (a) and (b) shows the lower bound on the reconstruction error for object 'Knight' on the left and 'Helmet' on the right.*

5.6.4 Synthetic Experiments

To further evaluate the proposed 3D method, we generate a dataset of two synthetic objects with spatially-varying material maps and challenging geometries. In the following, we describe our dataset generation, ground truth, and optimization processes before evaluating and presenting the results of our method on this synthetic data.

Dataset Creation: Starting from an initial 3D mesh with normal and (hand-designed) material information, we distribute 500 views uniformly on a sphere centered at the object and with radius samples uniformly in [0.8m, 1.2m]. For each view, we first project all parameters of this initial ground truth mesh into this view to get 2.5D parameter maps that we then render into the observation image using the forward renderer of our model. For that, we simulate our real sensor in terms of light and camera parameters and the proposed svBRDF model. When rendering the input mesh naively, the normal map would be piece-wise constant as each pixel of each face triangle would be assigned the same normal vector. This artifact stems from the limited mesh resolution. Thus, for a more realistic appearance, we do not calculate face normals from the mesh but, for each point on a face, interpolate the normals of the face's vertices – known as *shading normals* using Phong interpolation for the normals. Consequently, this shading normal differs from the geometrical normal for each point on a mesh's face. Our geometric consistency regularizer promotes normal and depth consistency in 2D by encouraging that the normal field integrates to the projected depth map. As this holds for geometrical normals but not shading normals, our synthetic normal and depth maps are not consistent in the sense of our geometric consistency regularizer. We will return to the question of ground-truth depth maps in the next paragraph.

	Initial GT Mesh		Fused GT Mesh	
	L_1	*RSME*	L_1	*RSME*
Knight	$1.06 \cdot 10^{-6}$	$8.38 \cdot 10^{-5}$	$9.29 \cdot 10^{-7}$	$9.84 \cdot 10^{-5}$
Helmet	$3.07 \cdot 10^{-6}$	$4.44 \cdot 10^{-5}$	$2.84 \cdot 10^{-6}$	$4.46 \cdot 10^{-5}$

Table 5.4: Quantitative Evaluation of Synthetic Experiments (3D). *For the objects 'Knight' and 'Helmet', we evaluate the final reconstruction of the proposed 3D method on rendered predictions of nine held-out test views. Shown are the averaged L_1 and RMSE losses, with the observation and prediction values ranging in [0, 1].*

Further, we note that it is important for the proposed optimization that both real and synthetic observations are in the same value range. Therefore, we model over-exposure of natural images by clamping the synthetic observations to the dynamic range of our real camera. In total, the synthetic dataset consists of (1) the rendered synthetic observation images, the poses, and the facetted, rough depth maps, which serve as input to the proposed 3D method, and (2) the rendered parameter maps for materials and geometry which serve as *initial ground truth*.

Ground Truth: By the design of our method, the mesh that is predicted by our method will not match the initial 3D mesh used for rendering the synthetic observations. In most cases, the vertex count will differ noticeably since our mesh fusion step uses a fixed voxel grid size of $0.5mm$ while the initial meshes can be arbitrary. This means that some intrinsic error between the synthetic observations and the final rendered predictions is to be expected, yielding a lower bound on the attainable reconstruction error. To quantify this error/lower bound and account for it during evaluation, we denote the initial mesh and its rendered observations and parameter maps as *initial ground truth* and additionally calculate the *fused ground truth* as follows: Using the proposed mesh fusion, we fuse these initial ground truth parameter maps back into a 3D mesh. Then, we render observations and project parameter maps from the resulting fused ground truth mesh into the same views as before. The differences between these two synthetic observations are shown in Fig. 5.13. For evaluation, we calculate the loss with respect to rendered observations of both ground truth meshes.

Optimization: We run the proposed method on the synthetic dataset as on the real data. The only difference is that we adapt the loss weights for the material smoothness and geometric consistency as follows: For iterations $1 - T/2$, we set $w_{\mathcal{GC}} = 1e5$, $w_{\mathcal{M}} = 5e3$ and for iterations $T/2 - T$ we increase $w_{\mathcal{GC}} = 1e5$, $w_{\mathcal{M}} = 3e4$. We use $|K| = 24$ keyframes and $m = 25$ neighboring observation views.

Evaluation:

We calculate the L_1 loss and the *Root Mean Squared Error* of the predictions with respect to both (1) the initial synthetic observations, which our loss optimizes for, and (2) the renderings from the fused ground truth mesh which encode the minimal achievable error. For better comparability, all material parameters are scaled to range [0, 1] before error calculation. We present results quantitatively in Tab. 5.4 and qualitatively in Fig. 5.14 and

Fig. 5.15. The results show that all parameters are reconstructed accurately. Only, and as expected, the material parameter roughness is unconstraint for pixels with a very low specular albedo value, e.g., the horse's legs of the object 'Knight' in Fig. 5.14. This verifies that the proposed 3D method recovers the ground truth from a set of multi-view images.

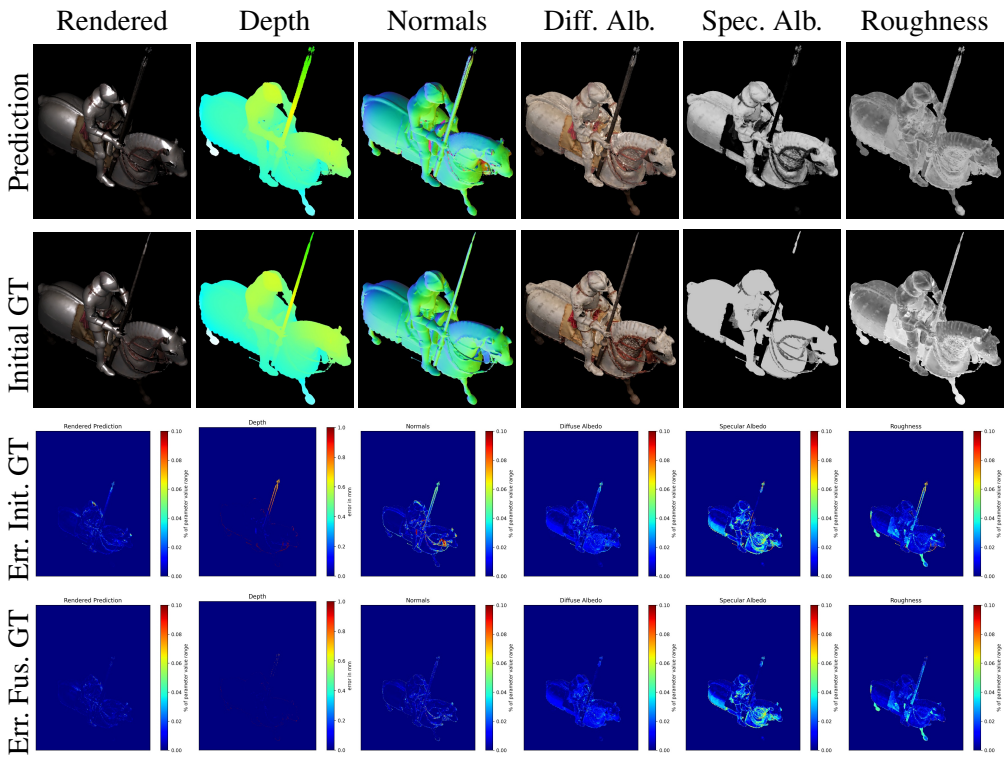


Figure 5.14: Synthetic Experiments 'Knight' (3D). *We show the final rendered predictions as well as rendered parameter maps for geometry and materials. For each parameter, we present our method's prediction, the initial ground truth (as in Fig. 5.13), and error maps with respect to the initial ground truth (the input observation) as well as the fused ground truth (the minimal attainable reconstruction error). We see that ground truth parameters are reconstructed well, the error for all parameters is* $< 0.05\%$ *of the respective maximum parameter value and in the order of sub-millimeters for the depth map. Note that for the object 'Knight', the specular albedo values of the horse's legs and the saddle are almost zero. This causes the respective roughness value being mostly unconstraint. Therefore, the errors in these pixels of the roughness map are expected.*

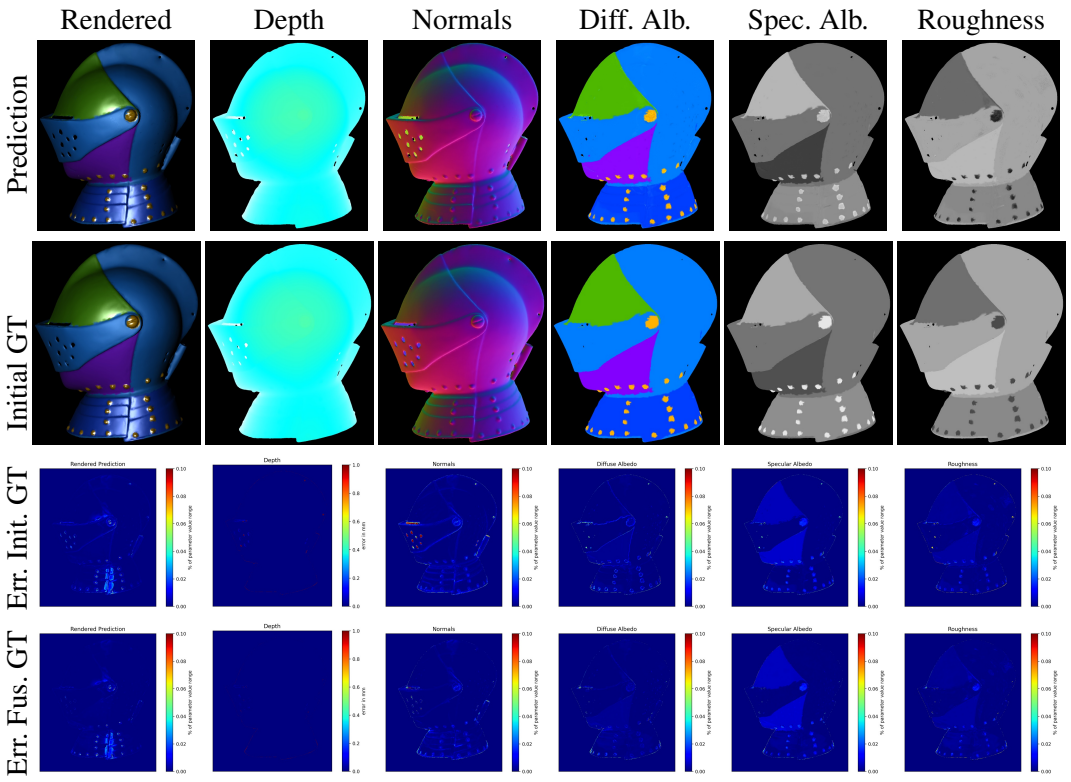


Figure 5.15: Synthetic Experiments 'Helmet' (3D). *We show the final rendered predictions as well as rendered parameter maps for geometry and materials. For each parameter, we present our method's prediction, the initial ground truth (as in Fig. 5.13), and error maps with respect to the initial ground truth (the input observation) as well as the fused ground truth (the minimal attainable reconstruction error). We see that ground truth parameters are reconstructed well, the error for all parameters is* $< 0.05\%$ *of the respective maximum parameter value and in the order of sub-millimeters for the depth map.*

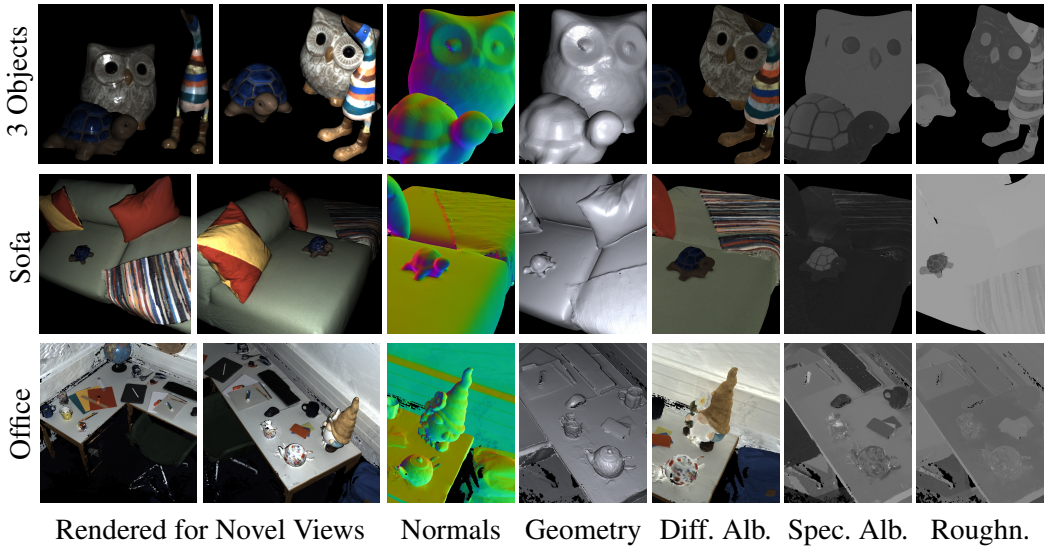


Figure 5.16: Reconstruction Beyond Object-Scale (3D). *We present reconstructions for three challenging scenes that are composed of multiple objects with non-convex shapes, detailed geometries, various different materials, many occlusions and shadows, and a spatial size of up to 2×3m, e.g., the scene 'Office'. Our method reconstructs detailed geometry and accurate materials leading to realistic renderings under new illumination and unseen viewpoints. Plus, we observe a clean separation of illumination effects and geometric properties, as the material maps are homogeneous per object parts (see, e.g., the 'Turtle' in '3 Objects' and 'Sofa').*

5.6.5 Reconstruction Results

We show the results of our method on captured objects in Fig. 5.17 and Fig. 5.18. They demonstrate the capabilities of our method to reconstruct accurate geometry and materials for a variety of real objects, scenes, and materials. Please see videos of our reconstructions here: https://sites.google.com/view/material-fusion/

Towards Scalable Scene Reconstructions: We demonstrate the scalability of the proposed 3D approach in Fig. 5.16. As shown, our method reconstructs scenes on the scale of several meters at a resolution of $\leq$ 2mm and recovers accurate appearance and geometry, leading to realistic renderings of novel viewpoints and illumination. Yet, the 'Office' scene shows two limitations of our model as we do not recover missing geometry or model global illumination effects. We include a discussion in the next Sec. 5.6.6.

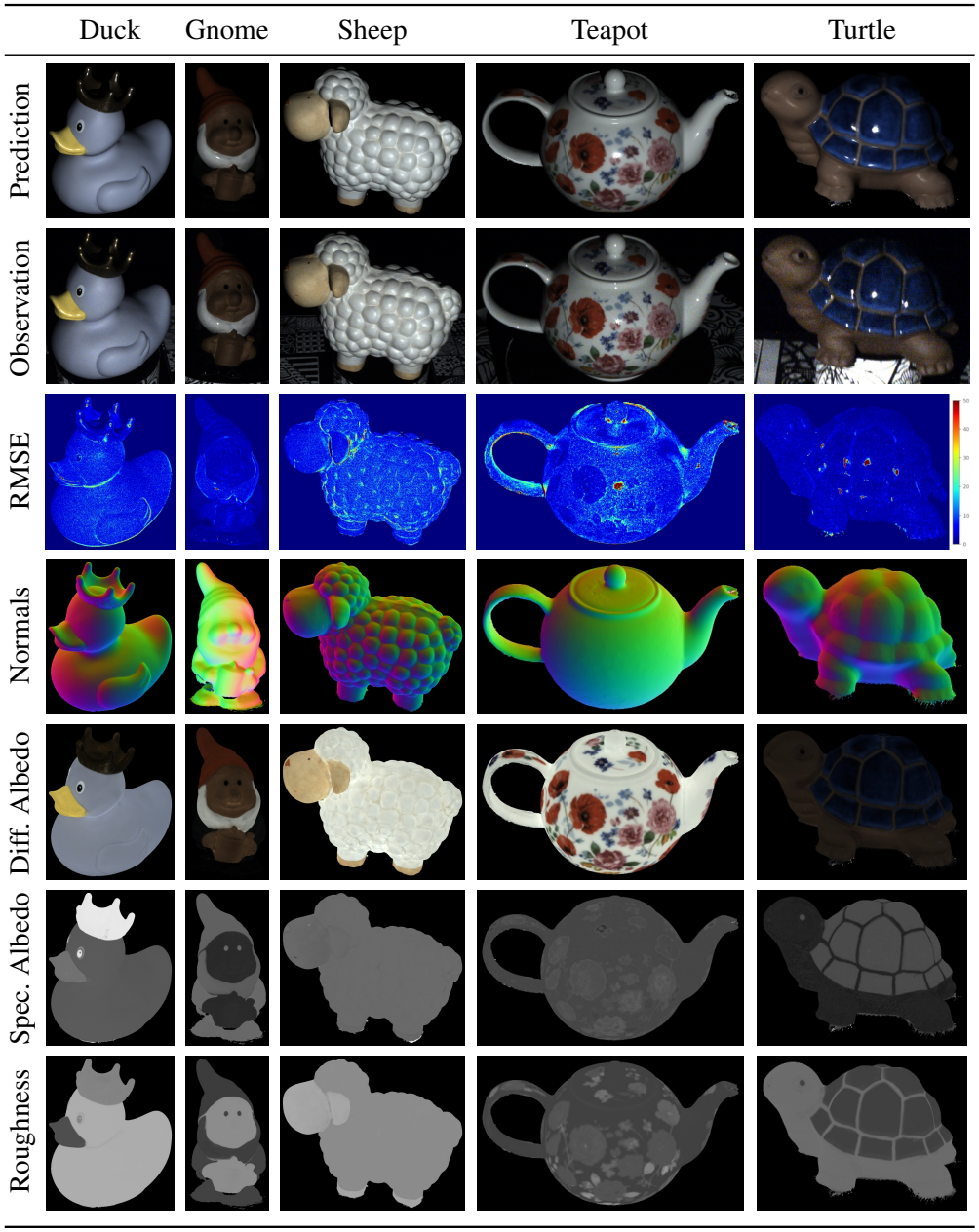


Figure 5.17: Reconstruction Results (3D) *for the objects 'Duck', 'Gnome', 'Sheep', 'Teapot', and 'Turtle'.*

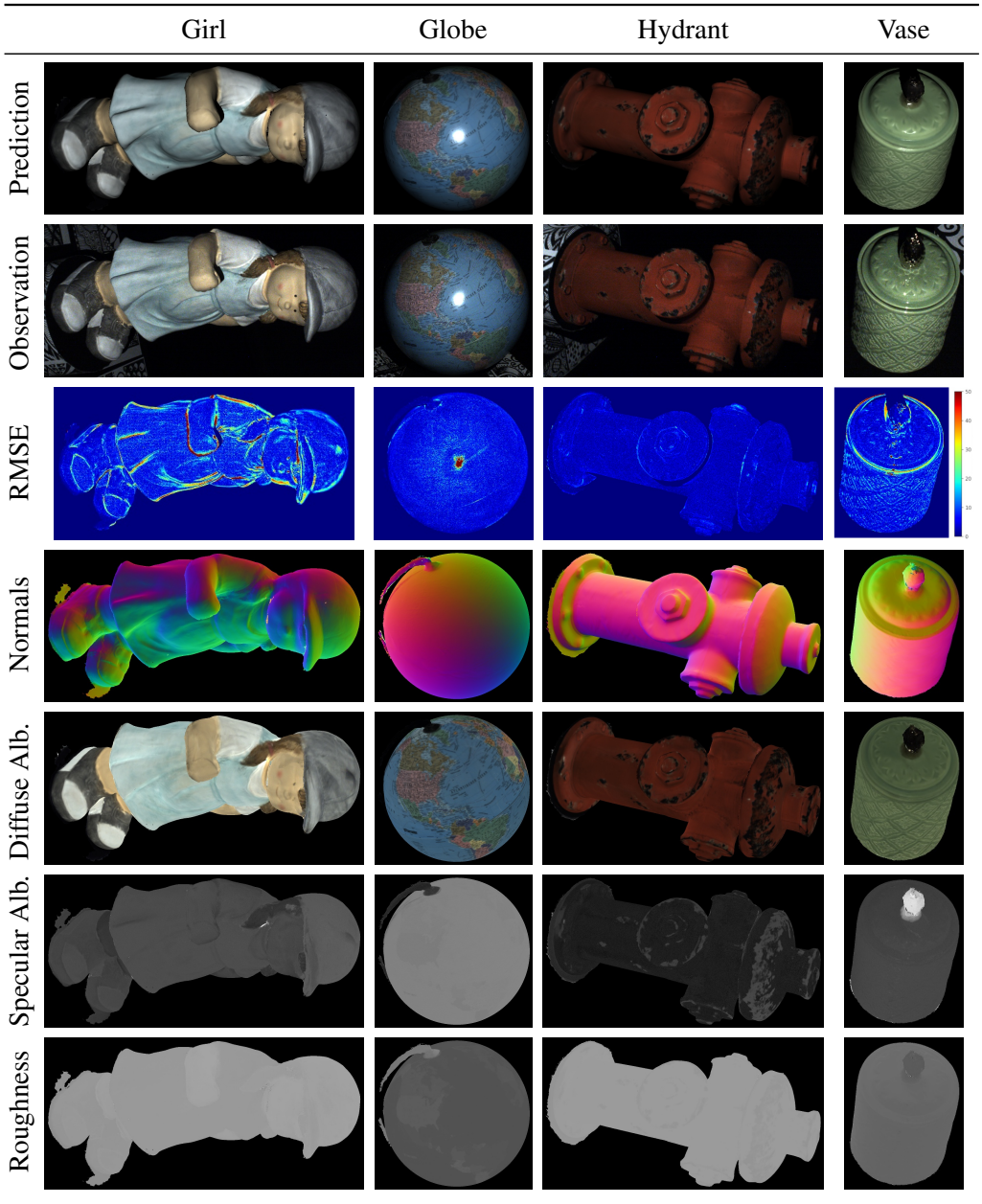


Figure 5.18: Reconstruction Results (3D) *for the objects 'Girl', 'Globe', 'Hydrant', and 'Vase'.*

5.6.6 Limitations and Outlook

The proposed 3D method refines depth maps but does not complete any missing geometry. Therefore, larger holes in the initial geometry can not be filled; see, e.g., the wall in the 'Office' scene in Fig. 5.16. Further, we decided on a renderer that models a single light bounce to keep computation tractable. That means global illumination, mirror-like materials, and strong inter-reflections cannot be modeled. Mild inter-reflections cause small local errors in the material maps, as visible in Fig. 5.17 for the 'Duck'. Last, the expressiveness of our BRDF model is limited. While it is able to represent most objects common in indoor rooms, it does not support anisotropic reflections or subsurface scattering.

To sum up, we have proposed a practical approach to estimating geometry and materials from a handheld sensor for full 3D models. Towards large-scale appearance and geometry reconstructions, we represented the scene as 2.5D parameter maps for a set of keyframes and introduced a distributed optimization scheme. Therefore, the processing memory requirements are almost independent of the scene size. This enables the method to process large numbers of input observations and optimization parameters, as demonstrated for multiple scenes with larger compositions of multiple objects. Further, we demonstrated that multi-view consistency regularization is key for accurate integration to a globally consistent 3D reconstruction. We hope this sparks future research towards scalable reconstruction of pose, geometry, and materials from handheld data.

We discuss both methods presented in this dissertation, remaining challenges, and future directions in the upcoming Chap. 6.

6 Discussion

While developing models for scalable multi-view reconstruction of geometry and materials from handheld data, we faced various challenges and found several solutions. Looking back, we identify four significant challenges that we consider crucial in the context of accurate geometry and reflectance estimation for indoor scenes. In this chapter, we elaborate on these, relate and discuss the work presented in this dissertation, include lessons learned, and look ahead.

The Inverse Rendering Problem: Naive optimization of the inverse rendering problem most likely gets stuck in local minima far from the optimal solution due to the high non-convexity of the optimization problem. In Sec. 6.1, we address this difficulty and explore causes, solutions, and possible future developments. Our focus is on creating stable optimization routines, effective regularization, and concise parameter representations.

Interlinked Properties: Geometry and material properties are strongly interlinked; separating them is key for reconstruction results that generalize to novel views, different lighting, or altered BRDFs. We relate our approach of optimizing all parameters jointly and look ahead in Sec. 6.2. Further, we discuss the lack of large-scale training data of real indoor scenes and elaborate on our method's potential to foster the creation of a dataset with pseudo ground truth for svBRDFs and geometry.

Handheld Capture Setup: Only data captured by handheld systems is practical for reconstructing arbitrary indoor environments. However, such capture data is inherently sparse and inhomogeneous with respect to BRDF samples. We discuss the challenges of 3D reconstruction from handheld data and contextualize our novel regularization approach and custom-built acquisition system in Sec. 6.3.

Scalability: Scaling 3D reconstruction beyond single objects usually encompasses more universal problem settings - hereby, raising diverse sets of new challenges. We elaborate on the resulting research questions, identify the common challenges shared between most tasks, and discuss our solutions and possible extensions in Sec. 6.4.

6.1 The Inverse Rendering Problem

Rendering describes methods to synthesize 2D images from a parametric scene description. Inverse rendering is the problem of inferring scene parameters from 2D images by differentiating rendering algorithms. The main objective is to find parameters that minimize the photometric error between rendered synthetic and real images. Hereby, automatic differentiation provides the derivatives of the loss wrt. parameters. But, as Nicolet et al. [NJJ21] nicely phrased, '*the availability of derivatives is no panacea*'. Or Vicini et al. [VSJ22] stated, '*While rendering is intrinsically differentiable, it can be difficult to retain this property in the transition from equation to algorithm.*' Even though gradients are available, optimization through inverse rendering is a challenging problem in itself. Due to a highly non-convex objective, naive approaches easily converge to poor solutions represented by local minima.

The plausibility and accuracy of the reconstructed scene parameters, as well as the stability and convergence of the optimization, are influenced by various properties, starting with the non-convexity of the problem, Sec. 6.1.1, over the captured input data, Sec. 6.1.2, to the optimization objective, Sec. 6.1.3, and the details of the optimization routine, Sec. 6.1.4.

In this section, we elaborate on all these aspects and discuss our proposed solutions and possible challenges of future developments. We also include our insights and lessons learned.

6.1.1 Non-Convexity

Scenes are composed of multiple, usually non-convex, objects. Naturally, (self-)occlusions and (self-)shadows occur, which are instantaneous changes on a pixel level. These cause sharp image visibility edges and introduce discontinuities to the rendering function. For differentiation, we calculate gradients of the image wrt. scene parameters. Now, if this includes a parameter that discontinuities depend on, e.g., the scene geometry, it biases gradient calculation as shown by Vicini et al. [VSJ22]. During optimization, these parameter-dependent discontinuities shift, yielding an inherently non-convex optimization landscape.

Additionally, observations can be explained in many ways: A pixel showing a 3D point can appear brighter due to a brighter diffuse albedo value, a more glossy BRDF, a different surface orientation, or a closer distance to the camera. The parameters for reflection and geometry are strongly correlated, as discussed in Sec. 6.2.

A third factor adding complexity to the inverse rendering problem is ambiguities: Many BRDF models exhibit some redundancy in return for, e.g., interpretability and expressiveness. Even a hypothetical non-redundant representation is under-constrained if the observation samples do not contain specular BRDF information. Thus, some ambiguities in the problem formulation are unavoidable.

Foremost, multi-view images and concise yet expressive parameter representations help to stabilize the optimization despite these discontinuities, correlations, and ambiguities. Our multi-view consistency term, as presented in Sec. 5.3.2, in combination with the parallel and distributed optimization scheme, see Sec. 5.4, effectively enable neighbor view information to propagate over the large scene parts and guide the optimization to prevent over-fitting to local minima, see Fig. 5.5. Further, we observe that compact and non-

ambiguous representations (e.g., quaternions for poses and deviation angles for normals) that fit the capture conditions (e.g., Fresnel term $F = 1$, as our sensor rig does not capture it) prove best. Our experiments confirm stable optimization convergence and consistent predictions across different keyframes.

Judged by the prominence in the latest publications, e.g., [Yao+22; JP22; WJP21; Sun+23b; Sun+23a; Wan+23; Zha+23c; Mai+23; Li+21a; Li+20b; Zhu+22a; Li+22c; Zhu+23a; YN23; Mao+23; Fan+23; Zha+23a; Wu+23a], generalization to unknown, natural, ambient, or indirect lighting conditions will be essential in future developments. Estimating lighting amplifies all the above factors greatly and also adds to them: For one, while capturing, the operator and sensor will interfere with the propagation of light rays in the scene. Second, the resolution to which, e.g., an environment map, can be reconstructed from partial observations, even theoretically, depends on the glossiness of the materials present in the scene. Last, as we usually observe lighting as reflections from surfaces or media only, lighting and materials are very interlinked (just as materials and geometry). These are all largely unsolved future challenges. However, it is clear that unknown lighting increases the number of light rays that need to be traced immensely. Thus, modeling complex light paths with multiple bounces will be indispensable. For that, highly optimized and fast differentiable renderers, like, e.g., Mitsuba 3 [Jak+22], will likely prove their worth quickly.

6.1.2 Real-World Capture Data

Similar to [Azi+19; Nim+21; Neh+05], we have observed that real-world captures exhibit unavoidable imperfections that can noticeably affect the quality of the reconstruction results. Such imperfections result from conversion inaccuracies and capture noise. Regular 2D images are the results of calculations on raw sensor data. Therefore, small inaccuracies or uncertainties in calibration, demosaicing, devignetting, or image undistortion add up and displace the pixel measurements. Next to the image conversion, capture noise is a source of imperfections: It stems from diffraction and aberration effects in lenses as well as the so-called *circle-of-confusion* on the sensor. Every traditional lens can only focus at a single distance and does not focus all rays perfectly, which shows as a blur spot on the sensor – the circle-of-confusion. It is proportional to the aperture and anti-proportional to the effective image resolution, i.e., the sharpness. For high quality cameras, the effective resolution is likely reduced by a factor of two for a half-open aperture. In our captures, we use a single, known light source to keep computation tractable and avoid the uncertainties of lighting estimation, see Sec. 6.1.1. This necessitates capturing in the dark, which prohibits setting both a small aperture and a short exposure time (even when using strong point light sources). Therefore, we compromise a small exposure and little motion blur for a small aperture and concede some loss in effective image resolution. To sum up: While in theory, there is one solution to the inverse rendering problem that depicts reality and minimizes the loss globally, it might not exist in practice due to this mixture of imperfections.

We have designed a sensor-specific, extensive, and rather laborious calibration pipeline to minimize such inaccuracies. Yet, our experience shows that capture conditions always vary when capturing outside of laboratory environments and at different locations. For example, we found that some physical component of our capture rig is temperature dependent. Or, for

the part of the calibration that was generally expected to be invariant to ambient lighting (daylight in our case), we noticed that our calibration results still differed slightly between seasons. These factors should best be eliminated, but it is neither universally possible nor feasible for changing capture locations. Thus, a robust algorithm that can handle some imperfections in the data is crucial for a versatile method.

Therefore, the proposed methods include pose adjustments in the joint optimization and account for uncertainties. E.g., since depth predictions are less confident for slanted surfaces, we use a soft regularizer instead of hard linking depth and normals. Further, as reflection information is not equally distributed across samples, see Sec. 6.3.2, we weigh pixels in our loss based on their information content. Our experiments confirm the effectiveness by showing sharp predictions for blurred input observations, Fig. 4.4, and coarse initial geometry, Fig. 4.9.

Thinking further, extending our method to, e.g., outdoors or large indoor areas will increase the differences in distances captured in each image. Depending on the *depth-of-field*, this will likely cause image regions to be out-of-focus. Also, when not captured in the dark, direct or indirect sunlight will further amplify the differences between light and dark image areas. Using some depth-of-field dependent pixel weighting and HDR images would be beneficial. However, handling HDR imagery poses new challenges, as it increases the bias of the reconstruction towards bright image regions. We will return to this point at the end of the next section.

6.1.3 Optimization Objective

When conducting experiments on inverse rendering, we regularly observe that slight decreases in the total loss need not result in qualitatively better renderings. While being close, the quantitative objective function does not model visual quality perfectly. We discuss this phenomenon and elaborate on the design of the optimization objective in the following.

Photometric Rendering Loss: There is an intrinsic discrepancy between the measured photometric loss and human perception. Since the human visual system's response to light is not linear, the same quantitative error looks different to a human's eye in bright and dark image regions. While this effect might be small in practice, human qualitative assessments of results naturally differ from computational quantitative measures (e.g., L_1 or L_2 loss).

Further, the common loss to infer scene parameters from RGB images is the photometric loss: For the same camera viewpoint, it measures pixel color differences between an image and the rendering of predicted scene parameters. By design, the optimal scene parameters will render into a predicted image that minimizes the loss. However, not all parameters have the same effect on the final rendering: Small pose perturbations can result in high errors as all pixels get shifted. And, as Nicolet et al. [NJJ21] stated, '*geometric distortion is generally invisible in renderings*' when reconstructing shape with a pure rendering loss. That necessitates additional regularization.

Regularization: Not all parameters are guided equally by the photometric loss, so plausible estimates must be promoted. Adding regularization terms to the optimization objective can help to migrate robustness issues (e.g., via smoothness terms), enforce physical constraints

(e.g., by penalizing depth-normal inconsistency), and improve predictions qualitatively (e.g., by the propagation of material properties). In Sec. 4.3.3, we introduce the bivariate material smoothness loss that effectively links specular reflectance properties of surface points based on their distance and diffuse albedo estimates. We show improved reconstructions of glossy highlights and more physically plausible parameter map estimates, see Fig. 4.7a. Luan et al. [Lua+21] also report positive results in follow-up work. In Sec. 5.3.2, we propose the multi-view consistency regularization to enable efficient optimizations locally while preserving global consistency. We demonstrate its necessity for realistic and accurate 3D models qualitatively and quantitatively, see Fig. 5.5.

These examples prove regularization to be necessary. Nevertheless, caution is advisable: Adding terms to the optimization objective forces the solution to compromise these terms, as noted in [Yao+22; NJJ21]. For one, this necessitates additional hyperparameters that must be chosen carefully, see Sec. 5.6.2. Second, it introduces discrepancies between our final evaluation metric (usually the photometric loss) and the optimization objective. These and the previously described difference between human perception and machine measurement explain the difficulty in finding a direct one-to-one correspondence between the optimization objective and qualitative human assessment, as stated at the beginning of this Sec. 6.1.3.

Range of Observations: The measured observations are the last aspect to consider in the context of the optimization objective. Our camera captures *Low Dynamic Range (LDR)* imagery with a saturation value $s \approx 2^{12}$, above which additional incoming light is not detected. Typical images fill this range $[0, s]$, as specular and non-specular intensity values often differ by more than an order of magnitude. The lower end of the range is characterized by an increase in sensor noise and a low signal-to-noise ratio. Closer to saturation, precision issues are likely: These high values of specular highlights are paired with small angles which, additionally, are taken to the power of four during BRDF calculation (see Eq. (4.2)). By experience, if numerical issues arise here, these propagate far and cause visible artifacts in the predicted renderings.

[Mat+03; NJR15; Xu+16] transform image observation values into log space to account for the intensity differences in HDR images. While this effectively counters the bias of the loss calculation towards bright image regions, for our LDR data, it would amplify dark image regions disproportionally. We observed the best qualitative results when (1) carefully accounting for numerical precision errors in code and (2) clipping our predictions to the saturation value of the observations, as this matches the visual range and perception of the observer best.

This section discussed the discrepancy between quantitative and qualitative results and examined the loss terms as well as the measured observations. Following this reasoning, experimenting with different loss terms that better model human visual perception is an exciting direction for future extensions.

6.1.4 Optimization Routine

In Sec. 5.4, we present a distributed optimization scheme that enables the optimization of many parameters and observations while staying within processing memory limits and allowing for effective regularization. The optimization is implemented using Pytorch's automatic differentiation and the off-the-shelf solver ADAM. Two details are important for stable optimization behavior:

First, we found that restricting the momentum of the ADAM optimizer helps convergence. A similar observation is noted in concurrent work by Nimier-David et al. [Nim+21]: '*noisy momentum is applied repeatedly to all variables [...] impeding convergence*'. Second, especially during early iterations, the optimization is prone to get stuck in local minima, also see [Bos+21a], due to, e.g., the ambiguities mentioned above in Sec. 6.1.1. To alleviate this problem, [Nim+21; Bos+22; Mai+17] apply a coarse-to-fine optimization scheme, [WWZ16; Mai+17] use a hierarchical approach, [Bos+20; Bos+21a] split the optimization before joint refinement, [YN23] transition from volume to surface rendering throughout the optimization, and [JP22] point out the importance of well-initialized parameter values. During the development of our method presented in Chap. 4, we employed a hierarchical optimization scheme but noticed it was not necessary for our final version, possibly due to our fully joint optimization approach. Instead, we leverage improved parameter initialization, as well as learning weight and learning rate adjustments over iterations, to bootstrap and guide the optimization of our method effectively. This way, we report stable optimization behavior and reliable convergence.

Recently, there have been new findings on the optimization routine: Nimier-David et al. [Nim+21] propose to sample observations in texture space instead of per-frame. Since we have not ablated different sampling strategies, this proposal is interesting for future development. Also, [VSJ22; BLD20; LHJ19] propose solutions to obtaining gradients at discontinuities wrt. shape parameters by reparameterizing the discontinuities for meshes or SDFs. We do not apply any method to handle, e.g., visibility discontinuities. Exploring potential applications of their methods for our approach will further benefit our optimization. Last, there are further options to explore for improving the parameter initialization. The latest publications on 3D reconstructions successfully use neural fields as building blocks in more extensive pipelines. E.g., [Zhu+22a; Sun+23b; Sun+23a; Wan+23] propose to use neural implicit fields with volume rendering as specialized sub-networks, e.g., to bootstrap their methods and then optimize the parameters of non-neural scene representations. Although potentially blurry, such a global 3D initialization of materials will likely stabilize the optimization behavior further.

6.2 Relation Geometry and Materials

We found the strong entanglement of geometry and materials to be one of the essential challenges for photo-realistic 3D reconstruction. The appearance captured in an image is the result of the orientation and distance of scene surfaces to the camera and the reflectance properties of each point on these surfaces. A pixel may appear brighter in an image due to any of these geometry or material-related factors. This strong correlation causes ambiguities in the inverse process of inferring geometry and materials from appearance in images. The methods that disentangle these properties well produce better reconstruction and generalization results. Our experiments show that by optimizing these correlated parameters jointly, a clean separation of parameters can be achieved - leading to accurate re-renderings and low generalization errors.

Instead of classical optimization, an alternative approach is training neural networks to predict parameter maps and learn the disentanglement from data. Here, we see a significant challenge in the training dataset. Creating reflectance data at scale for training is challenging and a huge effort. A versatile dataset featuring pseudo ground truth information would foster development significantly. Towards this goal, we proposed a method for practical and accurate geometry and reflectance estimation that can create pseudo ground truth from real-world scans. We hope to push future work toward creating such a dataset that will benefit research. In the following, we address disentanglement in Sec. 6.2.1 and dataset creation in Sec. 6.2.2, discuss our contribution, and look ahead.

6.2.1 Joint Optimization

To the best of our knowledge, we were the first to present a classical, optimization-based method that recovers geometry, materials, and poses by optimizing jointly instead of alternatingly. We found that the optimization strategy is important for fostering the separation of these strongly correlated parameters, facilitating generalization to more universal problem settings. Generally speaking, more constrained capture setups and parameter representations lead to fewer unknowns and easier-to-solve optimization problems. Early methods [Len+03; Li+20a; WZ15; WYT16; HS17; HS15; ZWT13; Tun+13] design distinct optimization stages and recover the shape before estimating the BRDF. Under very defined capture settings (e.g., in laboratory environments) or with restricted parameter representations, this allows for accurate reconstructions. [Gol+10; AZK08; Bir+06; Xia+16] show that more general capture settings, e.g., using a gantry arm or a tripod and a turntable, are feasible when alternating the optimization of geometric and photometric terms. They exploit that better geometry implies improved material estimates and vice versa. [Nam+18; GPG14] allow for capturing data of a handheld camera by carefully designing a complex optimization scheme that uses different objectives in an alternating optimization over shape and reflectance properties.

Our method, presented in Chap. 4, is the first to optimize jointly for geometry, poses, and svBRDFs using a unified objective function and input images from a freely moving capture system. The results in Sec. 4.6 show disentangled parameter maps and allow for novel view and relighting renderings. Our experiments further demonstrate that joint strategies enable

more accurate results compared to alternating optimization, see Fig. 4.10. Our finding on the positive impact of joint optimization is supported by subsequent works [Lua+21; VSJ22] and concurrent works on neural inverse rendering: [YN23; Bos+22; Bos+21a; Zha+21b; Bi+20b; Kan+19; Che+21b; Li+20b; Mao+23; Fan+23; Li+20b] point out the benefit of jointly solving for geometry and reflectance in network training.

The next steps will likely include modeling a wider variety of objects with more complex reflections, generalizing to natural or arbitrary lighting, and calculating higher-order illumination effects. Advances in these directions have been made by methods that assume well-known geometry, isolated objects, or estimate a single property: [LYC20] reconstruct transparent materials, [LNN23] estimate translucent objects, [Zhu+23b] approximate subsurface scattering, [Liu+23] reconstruct geometry and BRDF of very reflective objects with a single, highly glossy material in unknown environment, [Zha+23c] reconstruct volumetric media like smoke, [Guo+22; PHS20] compute mirror images, [Azi+19; JP22; Don+14] include natural illumination, [VSJ22; Nim+21] model multiple light bounces, [Zha+23a; Wu+23a] model inter-reflections, and [Had+23] approximate inverse global illumination. Developing algorithms that include such generalizations in unconstrained geometry and material reconstruction pipelines may impede parameter separation immensely. Likely, fostering disentanglement and creating sufficiently rich datasets should be addressed in future research. We elaborate on datasets in the next section.

6.2.2 Dataset Creation

As proven for many diverse tasks, neural networks can learn high-dimensional functions from correlated data. Therefore, they have the potential to learn the inverse image formation model and the complex interplay of materials, geometry, and lighting without requiring explicit formulations. Thus, they are essential in improving results and further generalizing problem formulations.

To discuss possible future development, we first review works on neural inverse rendering for inferring geometric and reflectance properties of indoor scenes. The following are all approaches presented for neural inverse rendering in Sec. 3.4.1 and neural scene representation in Sec. 3.4.2 with a (sub-)network (pre-)trained with supervision: [Li+22c; Des+18; Gao+19; ZK21; YN23; Bos+20; Bos+21b; Zha+21c; Lic+21; ALL20; Bi+20c; Kan+19; Kim+17; LSC18; DLG21; Li+20b; Li+22d; Zhu+22b; Wan+21b; LYC20; ZK22; Guo+23; Zha+23b; Pra+22; SLS23; Zhu+22a]. Notably, 15 of these 26 approaches propose a novel dataset for training. In total, 21 datasets are leveraged for training by these 26 methods. They all strive to create a large set of training images with maximum photorealism. And all consent on the gap between synthetic and real data being problematic, yet, only five of the above approaches train (fully or partially) on real data. Recall Sec. 2.1.2, real data that can serve as ground truth for reflectance estimation is very challenging and laborious to measure. Though, creating a sizeable synthetic dataset of indoor environments with photo-realistic appearance is also very ambitious. Therefore, the proposed datasets either simplify geometries, restrict materials, limit the sampling domain, pre-define (approximate) poses, or use partial datasets to pre-train sub-networks. As a result, these 15 datasets serve different problem formulations and vary greatly - none serves all purposes. Ideally, there

was a versatile dataset with real images plus material and geometry information that could serve as ground truth.

As presented in Sec. 5.6.5, our quantitative and qualitative results show that the material and geometry information recovered by our method allows for rendering images under novel viewpoints and lighting, which are almost indistinguishable from real photographs. That enables rendering an arbitrary number of real-like observations once a scene is scanned. The open-source codebases allow the integration of different sensor data, cameras, or material models.

By proposing our method, we encourage future research efforts in the creation of a dataset of real observations with inferred pseudo ground truth on materials, geometry, and poses. Many new scenes and observations could be created from a relatively small dataset of such scans, either by combining segmentation and recombination algorithms or by generative models. These scenes could be rendered with, e.g., environment lighting, novel camera models, or additional objects. Such a large-scale dataset with pseudo ground truth for geometry and materials is likely beneficial for many tasks, such as material estimation, geometry reconstruction, relighting, novel view synthesis of indoor scenes, or training embodied agents in very realistic environments.

6.3 Handheld Capture System

Practical, handheld acquisition systems are necessary for all scenes that do not fit into a laboratory setting, yet, these systems are limited in the samples they can provide. The reflectance function of any physically-based material model depends on the incoming light and outgoing camera directions and the surface point's normal. The captured samples must densely cover this function domain to fit reflectance purely from data accurately. Such dense samples are practically impossible to acquire with a handheld system, especially considering larger-scale spaces. Indeed, for mostly diffuse materials, the reflectance function is almost constant over large intervals of the domain, and different samples carry equal information. However, for glossy materials, a point's specular parameters can only be determined reliably if a highlight is visible in any of its observations. In this case, it is crucial which part of the domain is sampled.

This results in two challenges for material estimation from handheld systems: The reflectance function cannot be fit purely from data as the samples are too sparse, see Sec. 6.3.1, and reflectance information is not equally distributed over the samples, see Sec. 6.3.2. To solve the first problem, we propose a material regularizer that effectively shares sampled reflectance information across pixels, leading to clean and meaningful per-point material estimates. To tackle the second problem, we design a multi-sensor capture system that balances the size and weight of the device with the diversity of samples to maximize the reflectance information within the set of captured samples. We elaborate on both aspects in the following.

6.3.1 svBRDF Estimation from Sparse Samples

To handle sparse samples, carefully designed regularizers or constraints are necessary. While spatially-varying material estimation is crucial for photo-realistic results, recall Sec. 3.2.2, fitting a BRDF model per object point is impossible with small numbers of samples. Usually, this problem is tackled by constraining the BRDF model, regularizing scene points to share samples, or both. Hereby, a careful balance is important: Too little regularization shows up as strong noise in the material predictions. Yet, too much regularization results in little material variation and over-fitting to specular highlights - visible as highlights '*baked-in*' to the diffuse albedo and surface normal predictions, see the ablation in Sec. 5.6.2.

To constrain the BRDF, [Che+21a; Nim+21; Kim+17] model a single specular BRDF, [Yao+22; Zha+23a; Wu+23b] encourage materials to be Lambertian, [Zha+21b; Bos+20; Fan+23] approximate the BRDF by Spherical Gaussians, [Mao+23; Zha+23c] enforce latent space sparsity, and [Nam+18; GPG14; Lua+21; Bi+20b; Bi+20a; Che+21a; Gao+19; RPG16; AWL15] assume a co-located light/camera setup which requires fronto-parallel samples per scene material for reliable BRDF recovery. Most of these works are designed to handle only a few different materials and do not scale to multi-object scenes. Alternatively, [Bos+21b; Bos+22] densely compress measured BRDF data to low-dimensional representations - such a measured BRDF model poses an alternative to our parameter-based model and is an exciting option for testing in future developments.

To share reflectance information across scene points, [Bos+22; Kua+22; Zhu+23a;

Liu+23] employ regularization in the form of material smoothness losses, and [Len+03; Nam+18] introduce a set of base BRDFs. Such regularization poses a soft constraint on the optimization parameters that allows for more variation and makes it better suited for multi-object scenes.

In our first work, presented in Chap. 4, we cluster points into base materials and propose a new regularization term. Instead of enforcing smoothness between all neighboring pixels or points, as [Bos+22; Kua+22; Zhu+23a; Liu+23] proposed, our regularizer allows for sharp material changes by encouraging smoothness based on not only the 3D distance but also the diffuse albedo. As the idea of base materials matches reality well for many objects, we present convincing results for single objects, see Fig. 4.2. Practically, determining the number of base materials from data is cumbersome, and while we propose automatic model selection in Sec. 4.4.2, this does not scale well to larger multi-object scenes.

In our follow-up work, presented in Chap. 5, we discover that the previously proposed regularization term is sufficient (without using base materials) if paired with more multi-view information. It also integrates well into the proposed multi-view consistent optimization and scales to scenes without requiring knowledge of the number of materials in the scene. We demonstrate the effectiveness of our regularizer in Fig. 5.5: The reconstructed roughness and specular albedo maps are clean and detailed without artifacts of baked-in specularities. The re-lighting results in Fig. 5.17 further demonstrate that reconstructed highlights are sharp and realistic.

6.3.2 Capturing Informative Samples of the BRDF

Reflectance information is not distributed equally over the BRDF domain. By performing the principal component analysis (PCA) on a measured BRDF database, Nielsen et al. [NJR15] found that relatively, most informative samples are obtained under close light- and view directions. This finding supports approaches that use co-located setups of light and camera [Nam+18; GPG14; Lua+21; Bi+20b; Bi+20a; Che+21a; Gao+19; RPG16; AWL15], which only capture these relatively most informative samples. The advantages are clear: This setup allows for using minimal acquisition systems (e.g., a phone and its flash), constrains the BRDF (see previous Sec. 6.3.1), and simplifies the rendering equation drastically as it avoids any shadowing. But Nielsen et al. also emphasize that '*the reported sampling directions do not cover the full variability of BRDFs*'. For that, a much larger sample space is needed.

Other methods account for this non-constant distribution of reflectance information by assuming the approximate positioning of the camera during captures of single objects, [Lic+21; Bos+20; AWL15]. This promises both sufficiently informative samples and easy, handheld capture routines. However, it is not practical beyond object scale since it relates the camera position to the object's orientation.

Instead of restricting the sample space or scene scale, we design a sensor system that balances portability (i.e., weight, size, one-hand-use) and large sample space. Generally, capturing many different distances and angles between light, scene surfaces, and camera implies more variety and information in the set of samples but also a larger, and thereby heavier, sensor rig. Our system, shown in Fig. 4.3, is small enough to be carried around cluttered scenes and light enough to be pointed at scenes from all sides by a human operator.

By placing one light just above the RGB camera, we account for the co-located samples but additionally include surrounding lights at varying distances to capture maximally diverse samples. This way, we target diverse views and light samples given an affordable handheld acquisition system. For experimental evidence, please see Sec. 5.6.4. It demonstrates that the samples taken by our handheld system allow for the recovery of material parameters close to ground truth. Due to the lack of ground truth data, we conduct this experiment on synthetic renderings which mirror the real sensor as closely as possible.

Concurrent and follow-up works support our system design as they are deploying similar sensor systems [Li+20a; WYT16]. A benefit, yet also a limitation, of this design is the single point light used per image. While known and calibrated lighting eliminates error sources during optimization and makes computation more feasible, multiple lights or environment lighting would provide denser BRDF samples. Potentially, it comes at the cost of higher computational demand and new challenges for calibration. These are compelling directions for future work.

6.4 Scalability

When discussing scalable 3D reconstruction, one typically means scaling to a larger spatial extent of the scene while keeping a particular resolution fixed. So scalability is usually accompanied by more universal problem settings, like minimal scanning setups, data of whole buildings or public spaces, real-time constraints, or casual scans with fewer restrictions on, e.g., lighting or materials. To this date, a one-solves-all solution is not yet realistic. But what all these problem settings have in common is a larger number of observations and optimization parameters (as a direct result of the increased spatial extent of the scene) and the need to enforce consistency of the reconstruction (since most observations capture the scene only partially). With our approach, as presented in Chap. 5, we propose a computationally feasible solution to these central problems of scalability. In the following, we present various challenges in the context of scalable 3D reconstruction, see Sec. 6.4.1, and discuss solutions and extensions in Sec. 6.4.2.

6.4.1 Challenges for Scalable Indoor Reconstructions

Reconstructing scenes of larger spatial extent with fixed resolution (or potentially higher resolution) implies increased input observations and optimization parameters. As processing memory is limited, this necessitates scene representations and reconstruction algorithms that avoid a high memory footprint and keep processing computationally feasible, as discussed in Sec. 3.5.1. Additionally, since observations cover only parts of the scene, it is crucial to establish consistency within the reconstruction. Last, while objects allow for outside-in captures, scans of larger scenes are taken inside-out. This leads to fewer observations per scene point and, practically, to physical restrictions during capturing due to, e.g., possibly occluding scene parts and limited body size of the human operator. Coverage issues and incomplete scans are unavoidable results.

Thus, there are intrinsic challenges of scalability that are shared for most problem settings in the context of indoor 3D reconstructions: Many observation views and optimization parameters, as well as consistency and occlusions. These intrinsic problems often accompany task-dependent challenges or arise from more universal problem formulations. Examples are:

(a) Ambient illumination [Li+22c; Li+22b; Nim+21; Phi+21; Wu+22; Hae+21], reflective or mirror-like materials [Guo+22; Li+22c; Str+19], or multiple light bounces [Nim+21; Azi+19; Li+22b] if targeting casual scans with unknown lighting or a great diversity in materials,

(b) Scene completion [Dai+21; Jia+20a; Hua+17], geometric registration with little overlap [Zen+17; CDK20; YYH20], or single view reconstruction [Li+22c] in case of sparse or minimal scanning setups,

(c) Global registration including several rooms and levels [LD21; CZK15; Cha+17], scaling implicit neural representations [Pen+20], or incremental integration [Whe+12] for scaling to full apartments or multi-level buildings,

(d) Temporary occluders/transient objects [Lee+21; Jun+21; SF16; Sch+16] or crowd-sourced images/videos [Zho+21; Li+22a; Gle+16] when aiming for reconstruction of large, public indoor spaces, and

(e) Real-time, large scale 3D reconstruction [CBI13; Nie+13].

Most of these works tackle either geometry, BRDF, or pose reconstruction, as presented in Sec. 3.5, and only three of them allow for modeling accurate geometry as well as reflections: Li et al. [Li+22c] estimate geometry and BRDF for a single panorama in 2.5D, Philip et al. [Phi+21] blend estimated 2D albedo, mirror, and irradiance images (no full BRDFs) for view synthesis under novel illumination, and Wu et al. [Wu+22] reconstruct 3D albedo and reflection parameters alongside geometry. Only Wu et al. compute a 3D model of the scene; we include a discussion in Sec. 6.4.2.

6.4.2 Solutions and Outlook

By combining global consistency constraints with local optimizations in 2.5D, we presented a method for which memory requirements during processing do not scale with scene size. Here, we trade off time for scale but not memory or accuracy. That renders our method computationally feasible for larger scene sizes and many observation views, experimentally verified in Sec. 5.6.5. This way, we present solutions to two intrinsic challenges of scalability (recall Sec. 6.4.1): Many observation views and optimization parameters (using a memory-efficient scene representation and a corresponding optimization scheme) and global consistency (via effective regularization).

A thought on coverage: Our handheld capture setup gives us the best chances to capture diverse observations regarding viewing and lighting angles. Still, coverage and large occlusions are problematic. Although our method performs well even for coarse and blurry initial geometry, see Fig. 4.4 and Fig. 4.9, it does not do any scene completion but depends on the completeness of the initial geometry. Note that it can easily run on top of depth estimation models that yield more complete and accurate results. Therefore, we consider the question of initializing the geometry as an orthogonal path forward. Our model can trivially benefit from these advances, and we do not explicitly target coverage issues in this dissertation.

Recent concurrent work by Wu et al. [Wu+22] propose a conceptually similar approach to ours for tackling scalability: They partition 3D space into tiles and fit two MLPs to each tile to encode view-independent and view-dependent appearance. All parameters of these local neural scene representations are stored in an octree, and the authors show great results for indoor scenes with diverse geometries and materials. Like our keyframe-based scene representation, their tile-based scene representation allows for parallel optimization, equally avoiding a high memory footprint. While we partition the set of input observations into local 2.5D keyframe representations, they partition 3D space into 3D tiles. Hereby, our model enables accurate reconstructions through the multi-view consistency loss, and similarly, their model depends on background sampling and reflection tile groups for globally consistent geometry and reflection estimates. Therefore, both models offer solutions to intrinsic challenges of scalability. A difference between the models lies in the material representations: By modeling view-dependent appearance via virtual lights underneath the surfaces, they

implicitly assume all glossy surfaces to be planar. Instead, we estimate svBRDF parameters for all surfaces of any shape.

Looking ahead, we see various possible extensions of the work presented in this dissertation as it could be combined with many of the tasks described in Sec. 6.4.1. Personally, we are most curious about including ambient illumination and more diverse material models (e.g., including metal or glass) to enable casual captures of arbitrary indoor scenes. This is very challenging, as it implies many more scene parameters. Simultaneously, it also implies a much less controlled environment due to an enormous increase in light paths contributing to each scene point's appearance.

In such a severely under-constrained setting, our current approach would require impossible numbers of measured observation samples. Moreover, more than multi-view information is likely required to resolve all ambiguities inherent in the interplay of geometry, materials, and lighting. Instead, combining our scalable model with neural representations to make up for this lack of information is the most promising. We hope our scalable model will foster research on such.

Abbreviations

BRDF	Bidirectional Reflectance Distribution Function
BTDF	Bidirectional Transmission Distribution Function
DOE	Diffractive Optical Element
GTR	Generalized-Trowbridge-Reitz
HDR	High Dynamic Range
IBR	Image Based Rendering
IR	Infrared
LDR	Low Dynamic Range
LED	Light-Emitting Diode
MC	Monte-Carlo
MERL database	Mitsubishi Electric Research Laboratories BRDF database
MLP	Multi-Layer Perceptron
MVPS	Multi-View Photometric Stereo
MVS	Multi-View Stereo
NeRF	Neural Radiance Fields
PDF	Probability Density Function
PS	Photometric Stereo
RANSAC	Random sample consensus
RGB	Red, Green, and Blue
RGB-D	Red, Green, Blue, and Depth
SDF	Signed Distance Function
SfM	Structure from Motion
SfS	Shape from Shading
svBRDF	spatially-varying Bidirectional Reflectance Distribution Function
TSDF	Truncated Signed Distance Function

Bibliography

[ABD10] A. Adams, J. Baek, and M. A. Davis. "Fast High-Dimensional Filtering Using the Permutohedral Lattice". In: *Computer Graphics Forum* 29.2 (2010), pp. 753–762. DOI: 10.1111/j.1467-8659.2009.01645.x.

[Aga+09] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski. "Building Rome in a Day". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2009. DOI: 10.1109/ICCV.2009.5459148.

[Alb+18] R. A. Albert, D. Y. Chan, D. B. Goldman, and J. F. O'Brien. "Approximate svBRDF Estimation From Mobile Phone Video". In: *Eurographics Symposium on Rendering - Experimental Ideas & Implementations*. 2018. DOI: 10.2312/sre.20181168.

[ALL20] L.-P. Asselin, D. Laurendeau, and J.-F. Lalonde. "Deep SVBRDF Estimation on Real Materials". In: *International Conference on 3D Vision*. 2020. DOI: 10.1109/3DV50981.2020.00126.

[ASH22] M. Asthana, W. Smith, and P. Huber. "Neural Apparent BRDF Fields for Multiview Photometric Stereo". In: *Proceedings of the ACM SIGGRAPH European Conference on Visual Media Production*. 2022. DOI: 10.1145/3565516.3565517.

[Ass+20] M. Assran, A. Aytekin, H. Feyzmahdavian, M. Johansson, and M. Rabbat. "Advances in Asynchronous Parallel and Distributed Optimization". In: *arXiv.org* (2020). DOI: 10.48550/arXiv.2006.13838.

[AWL13] M. Aittala, T. Weyrich, and J. Lehtinen. "Practical SVBRDF capture in the frequency domain". In: *ACM Transactions on Graphics* 32.4 (2013), 110:1–110:12. DOI: 10.1145/2461912.2461978.

[AWL15] M. Aittala, T. Weyrich, and J. Lehtinen. "Two-shot SVBRDF capture for stationary materials". In: *ACM Transactions on Graphics* 34.4 (2015), 110:1–110:13. DOI: 10.1145/2766967.

[Azi+19] D. Azinović, T.-M. Li, A. Kaplanyan, and M. Nießner. "Inverse Path Tracing for Joint Material and Lighting Estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019. DOI: 10.1109/CVPR.2019.00255.

[Azi+22] D. Azinović, R. Martin-Brualla, D. B. Goldman, M. Nießner, and J. Thies. "Neural RGB-D Surface Reconstruction". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022. DOI: 10.1109/CVPR52688.2022.00619.

[AZK08] N. G. Alldrin, T. E. Zickler, and D. J. Kriegman. "Photometric stereo with non-parametric and spatially-varying reflectance". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2008. DOI: 10.1109/CVPR.2008.4587656.

[Bad+20] A. Badki, O. Gallo, J. Kautz, and P. Sen. "Meshlet Priors for 3D Mesh Reconstruction". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020. DOI: 10.1109/CVPR42600.2020.00292.

[Bar+21] J. T. Barron, B. Mildenhall, M. Tancik, P. Hedman, R. Martin-Brualla, and P. P. Srinivasan. "Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021. DOI: 10.1109/ICCV48922.2021.00580.

[Bar+22] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. Hedman. "Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022). DOI: 10.1109/CVPR52688.2022.00539.

[Bar78] H. Barrow. "Recovering intrinsic scene characteristics from images". In: *Computer Vision Systems* (1978), pp. 3–26.

[Bi+20a] S. Bi, Z. Xu, P. P. Srinivasan, B. Mildenhall, K. Sunkavalli, M. Hašan, Y. Hold-Geoffroy, D. J. Kriegman, and R. Ramamoorthi. "Neural Reflectance Fields for Appearance Acquisition". In: *arXiv.org* (2020). DOI: 10.48550/arXiv.2008.03824.

[Bi+20b] S. Bi, Z. Xu, K. Sunkavalli, M. Hašan, Y. Hold-Geoffroy, D. Kriegman, and R. Ramamoorthi. "Deep Reflectance Volumes: Relightable Reconstructions from Multi-View Photometric Images". In: *Proceedings of the European Conference on Computer Vision*. 2020. DOI: 10.1007/978-3-030-58580-8_18.

[Bi+20c] S. Bi, Z. Xu, K. Sunkavalli, D. Kriegman, and R. Ramamoorthi. "Deep 3D Capture: Geometry and Reflectance from Sparse Multi-View Images". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020. DOI: 10.1109/CVPR42600.2020.00600.

[Bir+06] N. Birkbeck, D. Cobzas, P. F. Sturm, and M. Jägersand. "Variational Shape and Reflectance Estimation Under Changing Light and Viewpoints". In: *Proceedings of the European Conference on Computer Vision*. 2006. DOI: 10.1007/11744023_42.

[BKR17] S. Bi, N. K. Kalantari, and R. Ramamoorthi. "Patch-Based Optimization for Image-Based Texture Mapping". In: *ACM Transactions on Graphics* 36.4 (2017), 106:1–106:11. DOI: 10.1145/3072959.3073610.

[BLD20] S. Bangaru, T.-M. Li, and F. Durand. "Unbiased Warped-Area Sampling for Differentiable Rendering". In: *ACM Transactions on Graphics* 39.6 (2020), 245:1–245:18. DOI: 10.1145/3414685.3417833.

[Bli77] J. F. Blinn. “Models of Light Reflection for Computer Synthesized Pictures”. In: *Proceedings of SIGGRAPH* 11.2 (1977), pp. 192–198. DOI: 10.1145/563858.563893.

[BM15] J. T. Barron and J. Malik. “Shape, Illumination, and Reflectance from Shading”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.8 (2015), pp. 1670–1687. DOI: 10.1109/TPAMI.2014.2377712.

[Bos+20] M. Boss, V. Jampani, K. Kim, H. P. Lensch, and J. Kautz. “Two-shot Spatially-varying BRDF and Shape Estimation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020. DOI: 10.1109/CVPR42600.2020.00404.

[Bos+21a] M. Boss, R. Braun, V. Jampani, J. T. Barron, C. Liu, and H. P. Lensch. “NeRD: Neural Reflectance Decomposition from Image Collections”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021. DOI: 10.1109/ICCV48922.2021.01245.

[Bos+21b] M. Boss, V. Jampani, R. Braun, C. Liu, J. T. Barron, and H. P. Lensch. “Neural-PIL: Neural Pre-Integrated Lighting for Reflectance Decomposition”. In: *Advances in Neural Information Processing Systems*. 2021. DOI: 10.48550/arXiv.2110.14373.

[Bos+22] M. Boss, A. Engelhardt, A. Kar, Y. Li, D. Sun, J. T. Barron, H. P. Lensch, and V. Jampani. “SAMURAI: Shape And Material from Unconstrained Real-world Arbitrary Image collections”. In: *Advances in Neural Information Processing Systems*. 2022. DOI: 10.48550/arXiv.2205.15768.

[Bur12] B. Burley. *Physically-Based Shading at Disney*. Tech. rep. Walt Disney Animation Studios, 2012.

[Can86] J. Canny. “A Computational Approach to Edge Detection”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 8.6 (1986), pp. 679–698. DOI: 10.1109/TPAMI.1986.4767851.

[CBI13] J. Chen, D. Bautembach, and S. Izadi. “Scalable real-time volumetric surface reconstruction”. In: *ACM Transactions on Graphics* 32.4 (2013), pp. 1–16. DOI: 10.1145/2461912.2461940.

[CBR11] M. K. Chandraker, J. Bai, and R. Ramamoorthi. “A theory of differential photometric stereo for unknown isotropic BRDFs”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2011. DOI: 10.1109/CVPR.2011.5995603.

[CDK20] C. Choy, W. Dong, and V. Koltun. “Deep Global Registration”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020. DOI: 10.1109/CVPR42600.2020.00259.

[Cha+17] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang. “Matterport3D: Learning from RGB-D Data in Indoor Environments”. In: *International Conference on 3D Vision*. 2017. DOI: 10.1109/3DV.2017.00081.

[Che+21a] Z. Cheng, H. Li, Y. Asano, Y. Zheng, and I. Sato. "Multi-view 3D Reconstruction of a Texture-less Smooth Surface of Unknown Generic Reflectance". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021. DOI: 10.1109/CVPR46437.2021.01596.

[Che+21b] Z. Cheng, H. Li, R. I. Hartley, Y. Zheng, and I. Sato. "One Ring to Rule Them All: a simple solution to multi-view 3D-Reconstruction of shapes with unknown BRDF via a small Recurrent ResNet". In: *arXiv.org* (2021). DOI: 10.48550/arXiv.2104.05014.

[CK13] Q. Chen and V. Koltun. "A Simple Model for Intrinsic Image Decomposition with Depth Cues". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2013. DOI: 10.1109/ICCV.2013.37.

[CL96] B. Curless and M. Levoy. "A Volumetric Method for Building Complex Models from Range Images". In: *ACM Transactions on Graphics*. 1996, pp. 303–312. DOI: 10.1145/237170.237269.

[CT82] R. L. Cook and K. E. Torrance. "A Reflectance Model for Computer Graphics". In: *ACM Transactions on Graphics* 1.1 (1982), pp. 7–24. DOI: 10.1145/800224.806819.

[CZK15] S. Choi, Q. Zhou, and V. Koltun. "Robust reconstruction of indoor scenes". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2015. DOI: 10.1109/CVPR.2015.7299195.

[Dai+17] A. Dai, M. Nießner, M. Zollöfer, S. Izadi, and C. Theobalt. "BundleFusion: Real-time Globally Consistent 3D Reconstruction using On-the-fly Surface Re-integration". In: 2017. DOI: 10.1145/3516521.

[Dai+21] A. Dai, Y. Siddiqui, J. Thies, J. Valentin, and M. Nießner. "SPSG: Self-Supervised Photometric Scene Generation from RGB-D Scans". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021. DOI: 10.1109/CVPR46437.2021.00179.

[Des+18] V. Deschaintre, M. Aittala, F. Durand, G. Drettakis, and A. Bousseau. "Single-image SVBRDF capture with a rendering-aware deep network". In: *ACM Transactions on Graphics* 37.4 (2018), pp. 1–15. DOI: 10.1145/3197517.3201378.

[DG19] S. Donné and A. Geiger. "Learning Non-volumetric Depth Fusion using Successive Reprojections". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019. DOI: 10.1109/CVPR.2019.00782.

[DJ18] J. Dupuy and W. Jakob. "An Adaptive Parameterization for Efficient Material Acquisition and Rendering". In: *ACM Transactions on Graphics* 37.6 (2018), 274:1–274:18. DOI: 10.1145/3272127.3275059.

[DLG21] V. Deschaintre, Y. Lin, and A. Ghosh. “Deep Polarization Imaging for 3D Shape and SVBRDF Acquisition”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021. DOI: 10.1109/CVPR46437.2021.01531.

[Don+14] Y. Dong, G. Chen, P. Peers, J. Zhang, and X. Tong. “Appearance-from-motion: recovering spatially varying surface reflectance under unknown lighting”. In: *ACM Transactions on Graphics* 33.6 (2014), 193:1–193:12. DOI: 10.1145/2661229.2661283.

[Esl+18] S. M. A. Eslami, D. J. Rezende, F. Besse, F. Viola, A. S. Morcos, M. Garnelo, A. Ruderman, A. A. Rusu, I. Danihelka, K. Gregor, D. P. Reichert, L. Buesing, T. Weber, O. Vinyals, D. Rosenbaum, N. C. Rabinowitz, H. King, C. Hillier, M. M. Botvinick, D. Wierstra, K. Kavukcuoglu, and D. Hassabis. “Neural scene representation and rendering”. In: *SCIENCE* 360.6394 (2018), pp. 1204–1210. DOI: 10.1126/science.aar6170.

[EVC08] C. H. Esteban, G. Vogiatzis, and R. Cipolla. “Multiview Photometric Stereo”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30.3 (2008), pp. 548–554. DOI: 10.1109/TPAMI.2007.70820.

[Fan+18] H. Fan, L. Qi, J. Dong, G. Li, and H. Yu. “Dynamic 3D Surface Reconstruction Using a Hand-Held Camera”. In: *IECON - Conference of the IEEE Industrial Electronics Society*. 2018. DOI: 10.1109/IECON.2018.8592826.

[Fan+23] Y. Fan, I. Skorokhodov, O. Voynov, S. Ignatyev, E. Burnaev, P. Wonka, and Y. Wang. “Factored-NeuS: Reconstructing Surfaces, Illumination, and Materials of Possibly Glossy Objects”. In: *arXiv.org* (2023). DOI: 10.48550/arXiv.2305.17929.

[FP10] Y. Furukawa and J. Ponce. “Accurate, Dense, and Robust Multi-View Stereopsis”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32.8 (2010), pp. 1362–1376. DOI: 10.1109/CVPR.2007.383246.

[GA18] R. Guy and M. Agopian. *Filament Renderer*. 2018. URL: https://google.github.io/filament/Filament.html.

[Gao+19] D. Gao, X. Li, Y. Dong, P. Peers, K. Xu, and X. Tong. “Deep Inverse Rendering for High-Resolution SVBRDF Estimation from an Arbitrary Number of Images”. In: *ACM Transactions on Graphics* 38.4 (2019), pp. 1–15. DOI: 10.1145/3306346.3323042.

[Gao+20] D. Gao, G. Chen, Y. Dong, P. Peers, K. Xu, and X. Tong. “Deferred neural lighting: free-viewpoint relighting from unstructured photographs”. In: *ACM Transactions on Graphics* 39.6 (2020), 258:1–258:15. DOI: 10.1145/3414685.3417767.

[GCS23] Y. Gao, Y.-P. Cao, and Y. Shan. “SurfelNeRF: Neural Surfel Radiance Fields for Online Photorealistic Reconstruction of Indoor Scenes”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023). DOI: 10.48550/arXiv.2304.08971.

[Geh+11] P. V. Gehler, C. Rother, M. Kiefel, L. Zhang, and B. Schölkopf. "Recovering Intrinsic Images with a Global Sparsity Prior on Reflectance". In: *Advances in Neural Information Processing Systems*. 2011. DOI: 10.5555/2986459.2986545.

[Gho+09] A. Ghosh, T. Chen, P. Peers, C. A. Wilson, and P. Debevec. "Estimating Specular Roughness and Anisotropy from Second Order Spherical Gradient Illumination". In: *Computer Graphics Forum* 28.4 (2009), pp. 1161–1170. DOI: 10.1111/j.1467-8659.2009.01493.x.

[Gle+16] C. Gleason, A. Guo, G. Laput, K. Kitani, and J. P. Bigham. "VizMap: Accessible Visual Information Through Crowdsourced Map Reconstruction". In: *Proceedings of the International ACM SIGACCESS Conference on Computers and Accessibility*. 2016. DOI: 10.1145/2982142.2982200.

[Gol+10] D. B. Goldman, B. Curless, A. Hertzmann, and S. M. Seitz. "Shape and Spatially-Varying BRDFs from Photometric Stereo". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32.6 (2010), pp. 1060–1071. DOI: 10.1109/TPAMI.2009.102.

[GPF10] D. Gallup, M. Pollefeys, and J.-M. Frahm. "3d reconstruction using an n-layer heightmap". In: *Proceedings of the DAGM conference on Pattern recognition*. 2010. DOI: 10.5555/1926258.1926260.

[GPG14] S. Georgoulis, M. Proesmans, and L. J. V. Gool. "Tackling Shapes and BRDFs Head-On". In: *International Conference on 3D Vision*. 2014. DOI: 10.1109/3DV.2014.81.

[Gro+09] R. Grosse, M. K. Johnson, E. H. Adelson, and W. T. Freeman. "Ground truth dataset and baseline evaluations for intrinsic image algorithms". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2009. DOI: 10.1109/ICCV.2009.5459428.

[Guo+22] Y.-C. Guo, D. Kang, L. Bao, Y. He, and S.-H. Zhang. "NeRFReN: Neural Radiance Fields With Reflections". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022. DOI: 10.1109/CVPR52688.2022.01786.

[Guo+23] J. Guo, S. Lai, Q. Tu, C. Tao, C. Zou, and Y. Guo. "Ultra-High Resolution SVBRDF Recovery from a Single Image". In: *ACM Transactions on Graphics* 42.3 (2023). DOI: 10.1145/3593798.

[Had+23] S. Hadadan, G. Lin, J. Novák, F. Rousselle, and M. Zwicker. "Inverse Global Illumination using a Neural Radiometric Prior". In: *ACM Transactions on Graphics* (2023). DOI: 10.1145/3588432.3591553.

[Hae+18] B. Haefner, Y. Quéau, T. Möllenhoff, and D. Cremers. "Fight Ill-Posedness With Ill-Posedness: Single-Shot Variational Depth Super-Resolution From Shading". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018. DOI: 10.1109/CVPR.2018.00025.

[Hae+21] B. Haefner, S. Green, A. Oursland, D. Andersen, M. Goesele, D. Cremers, R. Newcombe, and T. Whelan. “Recovering Real-World Reflectance Properties and Shading From HDR Imagery”. In: *International Conference on 3D Vision*. 2021. DOI: 10.1109/3DV53792.2021.00115.

[Ham17] E. Hammon. *PBR Diffuse Lighting for GGX+Smith Microsurfaces*. Game Developers Conference. 2017.

[Hed+16] P. Hedman, T. Ritschel, G. Drettakis, and G. Brostow. “Scalable Inside-Out Image-Based Rendering”. In: *ACM Transactions on Graphics* 35.6 (2016), 231:1–231:11. DOI: 10.1145/2980179.2982420.

[Hed+18] P. Hedman, J. Philip, T. Price, J.-M. Frahm, G. Drettakis, and G. Brostow. “Deep Blending for Free-viewpoint Image-based Rendering”. In: 37.6 (2018), 257:1–257:15. DOI: 10.1145/3272127.3275084.

[Hen+13] P. Henry, D. Fox, A. Bhowmik, and R. Mongia. “Patch Volumes: Segmentation-Based Consistent Mapping with RGB-D Cameras”. In: *International Conference on 3D Vision*. 2013, pp. 398–405. DOI: 10.1109/3DV.2013.59.

[Hig+09] T. Higo, Y. Matsushita, N. Joshi, and K. Ikeuchi. “A hand-held photometric stereo camera for 3-D modeling”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2009. DOI: 10.1109/ICCV.2009.5459331.

[HLZ10] M. Holroyd, J. Lawrence, and T. E. Zickler. “A coaxial optical scanner for synchronous acquisition of 3D geometry and surface reflectance”. In: *ACM Transactions on Graphics* 29.4 (2010), 99:1–99:12. DOI: 10.1145/1778765.1778836.

[Hol+08] M. Holroyd, J. Lawrence, G. Humphreys, and T. E. Zickler. “A photometric approach for estimating normals and tangents”. In: *ACM Transactions on Graphics* 27.5 (2008), 133:1–133:9. DOI: 10.1145/1457515.1409086.

[Hor70] B. K. Horn. *Shape from Shading: A Method for Obtaining the Shape of a Smooth Opaque Object from One View*. Tech. rep. 1970. DOI: 10.5555/888673.

[HS05] A. Hertzmann and S. M. Seitz. “Example-Based Photometric Stereo: Shape Reconstruction with General, Varying BRDFs”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27.8 (2005), pp. 1254–1264. DOI: 10.1109/TPAMI.2005.158.

[HS15] Z. Hui and A. C. Sankaranarayanan. “A Dictionary-based Approach for Estimating Shape and Spatially-Varying Reflectance”. In: *IEEE International Conference on Computational Photography*. 2015. DOI: 10.1109/ICCPHOT.2015.7168363.

[HS17] Z. Hui and A. C. Sankaranarayanan. “Shape and Spatially-Varying Reflectance Estimation from Virtual Exemplars”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.10 (2017), pp. 2060–2073. DOI: 10.1109/TPAMI.2016.2623613.

[HSW89] K. Hornik, M. Stinchcombe, and H. White. "Multilayer feedforward networks are universal approximators". In: *Neural Networks* 2.5 (1989), pp. 359–366. DOI: 10.1016/0893-6080(89)90020-8.

[Hu+20] B. Hu, J. Guo, Y. Chen, M. Li, and Y. Guo. "DeepBRDF: A Deep Representation for Manipulating Measured BRDF". In: *Computer Graphics Forum* 39.2 (2020), pp. 157–166. DOI: 10.1111/cgf.13920.

[Hua+17] J. Huang, A. Dai, L. Guibas, and M. Nießner. "3DLite: Towards Commodity 3D Scanning for Content Creation". In: *ACM Transactions on Graphics* 36.6 (2017), 203:1–203:14. DOI: 10.1145/3130800.3130824.

[Hua+20] J. Huang, J. Thies, A. Dai, A. Kundu, C. Jiang, L. J. Guibas, M. Niessner, and T. Funkhouser. "Adversarial Texture Optimization From RGB-D Scans". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020. DOI: 10.1109/CVPR42600.2020.00163.

[Hug+13] J. F. Hughes, A. van Dam, M. McGuire, D. F. Sklar, J. D. Foley, S. Feiner, and K. Akeley. *Computer Graphics: Principles and Practice*. 3rd ed. Addison-Wesley, 2013. ISBN: 978-0-321-39952-6. DOI: 10.5860/choice.51-2713.

[Hui+17] Z. Hui, K. Sunkavalli, J. Lee, S. Hadap, J. Wang, and A. C. Sankaranarayanan. "Reflectance Capture Using Univariate Sampling of BRDFs". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2017. DOI: 10.1109/ICCV.2017.573.

[IH81] K. Ikeuchi and B. K. P. Horn. "Numerical Shape from Shading and Occluding Boundaries". In: *Artificial Intelligence* 17.1-3 (1981), pp. 141–184. DOI: 10.1016/0004-3702(81)90023-0.

[ISS20] M. Innmann, J. Süßmuth, and M. Stamminger. "BRDF-Reconstruction in Photogrammetry Studio Setups". In: *IEEE Winter Conference on Applications of Computer Vision*. 2020. DOI: 10.1109/WACV45572.2020.9093320.

[Iza+11] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, and A. Fitzgibbon. "KinectFusion: Real-Time 3D Reconstruction and Interaction Using a Moving Depth Camera". In: *Proceedings of the ACM Symposium on User Interface Software and Technology*. 2011. DOI: 10.1145/2047196.2047270.

[Jak+22] W. Jakob, S. Speierer, N. Roussel, M. Nimier-David, D. Vicini, T. Zeltner, B. Nicolet, M. Crespo, V. Leroy, and Z. Zhang. *Mitsuba 3 renderer*. 2022. DOI: 10.5281/zenodo.7669127. URL: https://mitsuba-renderer.org.

[Jak10] W. Jakob. *Mitsuba Renderer*. 2010. URL: http://www.mitsuba-renderer.org.

[Jan+17] M. Janner, J. Wu, T. D. Kulkarni, I. Yildirim, and J. B. Tenenbaum. "Self-Supervised Intrinsic Image Decomposition". In: *Advances in Neural Information Processing Systems*. 2017. DOI: 10.5555/3295222.3295343.

[Jen+14] R. R. Jensen, A. L. Dahl, G. Vogiatzis, E. Tola, and H. Aanæs. “Large Scale Multi-view Stereopsis Evaluation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2014. DOI: 10.1109/CVPR.2014.59.

[Jia+20a] C. Jiang, A. Sud, A. Makadia, J. Huang, M. Nießner, and T. Funkhouser. “Local Implicit Grid Representations for 3D Scenes”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020. DOI: 10.1109/CVPR42600.2020.00604.

[Jia+20b] Y. Jiang, D. Ji, Z. Han, and M. Zwicker. “SDFDiff: Differentiable Rendering of Signed Distance Fields for 3D Shape Optimization”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020. DOI: 10.1109/CVPR42600.2020.00133.

[JK07] N. Joshi and D. J. Kriegman. “Shape from Varying Illumination and Viewpoint”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2007. DOI: 10.1109/ICCV.2007.4409021.

[JP22] A. Joy and C. Poullis. “Multi-view Gradient Consistency for SVBRDF Estimation of Complex Scenes under Natural Illumination”. In: *arXiv.org* (2022). DOI: 10.48550/ARXIV.2202.13017.

[Jun+21] D. Jung, J. Choi, Y. Lee, D. Kim, C. Kim, D. Manocha, and D. Lee. “DnD: Dense Depth Estimation in Crowded Dynamic Indoor Scenes”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021. DOI: 10.1109/ICCV48922.2021.01256.

[Kaj86] J. T. Kajiya. “The rendering equation”. In: *ACM Transactions on Graphics*. Vol. 20. 4. 1986, pp. 143–150. DOI: 10.1145/15886.15902.

[Kan+19] K. Kang, C. Xie, C. He, M. Yi, M. Gu, Z. Chen, K. Zhou, and H. Wu. “Learning Efficient Illumination Multiplexing for Joint Capture of Reflectance and Shape”. In: *ACM Transactions on Graphics* 38.6 (2019), 165:1–165:12. DOI: 10.1145/3355089.3356492.

[Kay+22] B. Kaya, S. Kumar, F. Sarno, V. Ferrari, and L. V. Gool. “Neural Radiance Fields Approach to Deep Multi-View Photometric Stereo”. In: *IEEE Winter Conference on Applications of Computer Vision*. 2022. DOI: 10.1109/WACV51458.2022.00402.

[Kay+23] B. Kaya, S. Kumar, C. Oliveira, V. Ferrari, and L. V. Gool. “Multi-View Photometric Stereo Revisited”. In: *IEEE Winter Conference on Applications of Computer Vision*. 2023. DOI: 10.1109/WACV56688.2023.00314.

[KB15] D. P. Kingma and J. Ba. “Adam: A Method for Stochastic Optimization”. In: *International Conference on Learning Representations*. 2015. DOI: 10.48550/arXiv.1412.6980.

[Kim+16] S. Kim, K. Park, K. Sohn, and S. Lin. "Unified Depth Prediction and Intrinsic Image Decomposition from a Single Image via Joint Convolutional Neural Fields". In: *Proceedings of the European Conference on Computer Vision*. 2016. DOI: 10.1007/978-3-319-46484-8_9.

[Kim+17] K. Kim, J. Gu, S. Tyree, P. Molchanov, M. Nießner, and J. Kautz. "A Lightweight Approach for On-the-Fly Reflectance Estimation". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2017. DOI: 10.1109/ICCV.2017.12.

[Kov+17] B. Kovacs, S. Bell, N. Snavely, and K. Bala. "Shading Annotations in the Wild". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2017. DOI: 10.1109/CVPR.2017.97.

[Kua+22] Z. Kuang, K. Olszewski, M. Chai, Z. Huang, P. Achlioptas, and S. Tulyakov. "NeROIC: Neural Rendering of Objects from Online Image Collections". In: *ACM Transactions on Graphics* 41.4 (2022), pp. 1–12. DOI: 10.1145/3528223.3530177.

[KUH18] H. Kato, Y. Ushiku, and T. Harada. "Neural 3D Mesh Renderer". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018. DOI: 10.1109/CVPR.2018.00411.

[KZ02] V. Kolmogorov and R. Zabih. "Multi-Camera Scene Reconstruction via Graph Cuts". In: *Proceedings of the European Conference on Computer Vision*. 2002. DOI: 10.1007/3-540-47977-5_6.

[Laf+12] P.-Y. Laffont, A. Bousseau, S. Paris, F. Durand, and G. Drettakis. "Coherent Intrinsic Images from Photo Collections". In: *ACM Transactions on Graphics* 31.6 (2012), pp. 1–11. DOI: 10.1145/2366145.2366221.

[Laf+13] F. Lafarge, R. Keriven, M. Bredif, and H.-H. Vu. "A Hybrid Multiview Stereo Algorithm for Modeling Urban Scenes". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.1 (2013), pp. 5–17. DOI: 10.1109/TPAMI.2012.84.

[Laf+97] E. P. F. Lafortune, S.-C. Foo, K. E. Torrance, and D. P. Greenberg. "Non-linear Approximation of Reflectance Functions". In: *Proceedings of SIGGRAPH*. 1997, pp. 117–126. DOI: 10.1145/258734.258801.

[Lam60] J. H. Lambert. *Photometria sive de mensura et gradibus luminis, colorum et umbrae*. Augustae Vindelicorum [Augsburg], Sumptibus Vidvae Eberhardi Klett, 1760. DOI: 10.3931/e-rara-9488.

[Law+06] J. Lawrence, A. Ben-Artzi, C. DeCoro, W. Matusik, H. Pfister, R. Ramamoorthi, and S. Rusinkiewicz. "Inverse Shade Trees for Non-Parametric Material Representation and Editing". In: *ACM Transactions on Graphics* 25.3 (2006), pp. 735–745. DOI: 10.1145/1141911.1141949.

[LB14] M. M. Loper and M. J. Black. "OpenDR: An Approximate Differentiable Renderer". In: *Proceedings of the European Conference on Computer Vision*. 2014. DOI: 10.1007/978-3-319-10584-0_11.

[LC87] W. E. Lorensen and H. E. Cline. “Marching Cubes: A High Resolution 3D Surface Construction Algorithm”. In: *ACM Transactions on Graphics*. 1987, pp. 163–169. DOI: 10.1145/37401.37422.

[LD21] G. Lim and N. Doh. “Automatic Reconstruction of Multi-Level Indoor Spaces from Point Cloud and Trajectory”. In: *Sensors* 21.10 (2021). DOI: 10.3390/s21103493.

[Lee+21] D. Lee, S. Ryu, S. Yeon, Y. Lee, D. Kim, C. Han, Y. Cabon, P. Weinzaepfel, G. Nicolas, G. Csurka, and M. Humenberger. “Large-scale Localization Datasets in Crowded Indoor Spaces”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021. DOI: 10.1109/CVPR46437.2021.00324.

[Len+03] H. P. A. Lensch, J. Kautz, M. Goesele, W. Heidrich, and H. Seidel. “Image-based reconstruction of spatial appearance and geometric detail”. In: *ACM Transactions on Graphics* 22.2 (2003), pp. 234–257. DOI: 10.1145/636886.636891.

[Lev+00] M. Levoy, K. Pulli, B. Curless, S. Rusinkiewicz, D. Koller, L. Pereira, M. Ginzton, S. Anderson, J. Davis, J. Ginsberg, J. Shade, and D. Fulk. “The Digital Michelangelo Project: 3D Scanning of Large Statues”. In: *ACM Transactions on Graphics*. 2000. DOI: 10.1145/344779.344849.

[LHJ19] G. Loubet, N. Holzschuch, and W. Jakob. “Reparameterizing discontinuous integrands for differentiable rendering”. In: *ACM Transactions on Graphics* 38.6 (2019). DOI: 10.1145/3355089.3356510.

[Li+18] T. Li, M. Aittala, F. Durand, and J. Lehtinen. “Differentiable Monte Carlo ray tracing through edge sampling”. In: *ACM Transactions on Graphics* 37.6 (2018), pp. 1–11. DOI: 10.1145/3272127.3275109.

[Li+20a] M. Li, Z. Zhou, Z. Wu, B. Shi, C. Diao, and P. Tan. “Multi-View Photometric Stereo: A Robust Solution and Benchmark Dataset for Spatially Varying Isotropic Materials”. In: *IEEE Transactions on Image Processing* 29 (2020), pp. 4159–4173. DOI: 10.1109/TIP.2020.2968818.

[Li+20b] Z. Li, M. Shafiei, R. Ramamoorthi, K. Sunkavalli, and M. Chandraker. “Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and svbrdf from a single image”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020. DOI: 10.1109/CVPR42600.2020.00255.

[Li+21a] R. Li, G. Zang, M. Qi, and W. Heidrich. “Shape and Reflectance Reconstruction in Uncontrolled Environments by Differentiable Rendering”. In: *arXiv.org* (2021). DOI: 10.48550/arXiv.2110.12975.

[Li+21b] Z. Li, T. Yu, S. Sang, S. Wang, M. Song, Y. Liu, Y.-Y. Yeh, R. Zhu, N. B. Gundavarapu, J. Shi, S. Bi, H.-X. Yu, Z. Xu, K. Sunkavalli, M. Havs.an, R. Ramamoorthi, and M. Chandraker. "OpenRooms: An Open Framework for Photorealistic Indoor Scene Datasets". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021. DOI: 10.1109/CVPR46437.2021.00711.

[Li+22a] C. Li, W. Chai, X. Yang, and Q. Li. "Crowdsourcing-Based Indoor Semantic Map Construction and Localization Using Graph Optimization". In: *Sensors* 22.16 (2022). DOI: 10.3390/s22166263.

[Li+22b] Z. Li, L. Wang, M. Cheng, C. Pan, and J. Yang. "Multi-view Inverse Rendering for Large-scale Real-world Indoor Scenes". In: *arXiv.org* (2022). DOI: 10.48550/arXiv.2211.10206.

[Li+22c] Z. Li, L. Wang, X. Huang, C. Pan, and J. Yang. "PhyIR: Physics-based Inverse Rendering for Panoramic Indoor Images". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022. DOI: 10.1109/CVPR52688.2022.01238.

[Li+22d] Z. Li, J. Shi, S. Bi, R. Zhu, K. Sunkavalli, M. Hašan, Z. Xu, R. Ramamoorthi, and M. Chandraker. "Physically-based Editing of Indoor Scene Lighting from a Single Image". In: *Proceedings of the European Conference on Computer Vision*. 2022. DOI: 10.1007/978-3-031-20068-7_32.

[Li+23] Z. Li, Q. Zheng, B. Shi, G. Pan, and X. Jiang. "DANI-Net: Uncalibrated Photometric Stereo by Differentiable Shadow Handling, Anisotropic Reflectance Modeling, and Neural Inverse Rendering". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023. DOI: 10.48550/arXiv.2303.15101.

[Lic+21] D. Lichy, J. Wu, S. Sengupta, and D. W. Jacobs. "Shape and Material Capture at Home". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021. DOI: 10.1109/CVPR46437.2021.00606.

[Lim+05] J. Lim, J. Ho, M. Yang, and D. J. Kriegman. "Passive Photometric Stereo from Motion". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2005. DOI: 10.1109/ICCV.2005.185.

[Liu+19] S. Liu, W. Chen, T. Li, and H. Li. "Soft Rasterizer: Differentiable Rendering for Unsupervised Single-View Mesh Reconstruction". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019. DOI: 10.1109/ICCV.2019.00780.

[Liu+20a] A. Liu, S. Ginosar, T. Zhou, A. A. Efros, and N. Snavely. "Learning to Factorize and Relight a City". In: *Proceedings of the European Conference on Computer Vision*. 2020. DOI: 10.1007/978-3-030-58548-8_32.

[Liu+20b] S. Liu, Y. Zhang, S. Peng, B. Shi, M. Pollefeys, and Z. Cui. “DIST: Rendering Deep Implicit Signed Distance Function with Differentiable Sphere Tracing”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020. DOI: 10.1109/CVPR42600.2020.00209.

[Liu+20c] Y. Liu, S. You, Y. Li, and F. Lu. “Unsupervised Learning for Intrinsic Image Decomposition From a Single Image”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020. DOI: 10.1109/CVPR42600.2020.00331.

[Liu+23] Y. Liu, P. Wang, C. Lin, X. Long, J. Wang, L. Liu, T. Komura, and W. Wang. “NeRO: Neural Geometry and BRDF Reconstruction of Reflective Objects from Multiview Images”. In: vol. 42. 4. 2023. DOI: 10.1145/3592134.

[LK11] S. Laine and T. Karras. “Efficient Sparse Voxel Octrees”. In: *IEEE Transactions on Visualization and Computer Graphics* 17.8 (2011), pp. 1048–1059. DOI: 10.1145/1730804.1730814.

[LM71] E. H. Land and J. J. McCann. “Lightness and Retinex Theory”. In: *Journal of the Optical Society of America* 61.1 (1971), pp. 1–11. DOI: 10.1364/JOSA.61.000001.

[LMC17] F. Logothetis, R. Mecca, and R. Cipolla. “Semi-Calibrated Near Field Photometric Stereo”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2017. DOI: 10.1109/CVPR.2017.481.

[LMC19] F. Logothetis, R. Mecca, and R. Cipolla. “A Differential Volumetric Approach to Multi-View Photometric Stereo”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019. DOI: 10.1109/ICCV.2019.00114.

[LN12] S. Lombardi and K. Nishino. “Single image multimaterial estimation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2012. DOI: 10.1109/CVPR.2012.6247681.

[LN16] S. Lombardi and K. Nishino. “Radiometric Scene Decomposition: Scene Reflectance, Illumination, and Geometry from RGB-D Images”. In: *International Conference on 3D Vision*. 2016. DOI: 10.1109/3DV.2016.39.

[LND18] C. Liu, S. G. Narasimhan, and A. W. Dubrawski. “Near-light photometric stereo using circularly placed point light sources”. In: *IEEE International Conference on Computational Photography*. 2018. DOI: 10.1109/ICCPHOT.2018.8368465.

[LNN23] C. Li, T. T. Ngo, and H. Nagahara. “Inverse Rendering of Translucent Objects using Physical and Neural Renderers”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023. DOI: 10.48550/arXiv.2305.08336.

[LS18a] Z. Li and N. Snavely. “CGIntrinsics: Better Intrinsic Image Decomposition through Physically-Based Rendering”. In: *Proceedings of the European Conference on Computer Vision*. 2018. DOI: 10.1007/978-3-030-01219-9_23.

[LS18b] Z. Li and N. Snavely. "Learning Intrinsic Image Decomposition from Watching the World". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018. DOI: 10.1109/CVPR.2018.00942.

[LSC18] Z. Li, K. Sunkavalli, and M. Chandraker. "Materials for Masses: SVBRDF Acquisition with a Single Mobile Phone Image". In: *Proceedings of the European Conference on Computer Vision*. 2018. DOI: 10.1007/978-3-030-01219-9_5.

[Lu+10] Z. Lu, Y. Tai, M. Ben-Ezra, and M. S. Brown. "A framework for ultra high resolution 3D imaging". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2010. DOI: 10.1109/CVPR.2010.5539829.

[Lua+21] F. Luan, S. Zhao, K. Bala, and Z. Dong. "Unified Shape and SVBRDF Recovery using Differentiable Monte Carlo Rendering". In: *Computer Graphics Forum* 40.4 (2021), pp. 101–113. DOI: 10.1111/cgf.14344.

[LYC20] Z. Li, Y.-Y. Yeh, and M. Chandraker. "Through the Looking Glass: Neural 3D Reconstruction of Transparent Shapes". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020. DOI: 10.1109/CVPR42600.2020.00134.

[Mai+17] R. Maier, K. Kim, D. Cremers, J. Kautz, and M. Nießner. "Intrinsic3D: High-Quality 3D Reconstruction by Joint Appearance and Geometry Optimization with Spatially-Varying Lighting". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2017. DOI: 10.1109/ICCV.2017.338.

[Mai+23] A. Mai, D. Verbin, F. Kuester, and S. Fridovich-Keil. "Neural Microfacet Fields for Inverse Rendering". In: *arXiv.org* (2023). DOI: 10.48550/arXiv.2303.17806.

[Mao+23] S. Mao, C. Wu, Z. Shen, and L. Zhang. "NeuS-PIR: Learning Relightable Neural Surface using Pre-Integrated Rendering". In: *arXiv.org* (2023). DOI: 10.48550/arXiv.2306.07632.

[Mat+03] W. Matusik, H. Pfister, M. Brand, and L. McMillan. "A Data-Driven Reflectance Model". In: *ACM Transactions on Graphics* 22.3 (2003), pp. 759–769. DOI: 10.1145/882262.882343.

[Mec+14a] R. Mecca, A. Tankus, A. Wetzler, and A. M. Bruckstein. "A Direct Differential Approach to Photometric Stereo with Perspective Viewing". In: *SIAM Journal on Imaging Sciences* 7.2 (2014), pp. 579–612. DOI: 10.1137/120902458.

[Mec+14b] R. Mecca, A. Wetzler, A. M. Bruckstein, and R. Kimmel. "Near Field Photometric Stereo with Point Light Sources". In: *SIAM Journal on Imaging Sciences* 7.4 (2014), pp. 2732–2770. DOI: 10.1137/140968100.

[Mec+16] R. Mecca, Y. Quéau, F. Logothetis, and R. Cipolla. "A Single-Lobe Photometric Stereo Approach for Heterogeneous Material". In: *SIAM Journal on Imaging Sciences* 9.4 (2016), pp. 1858–1888. DOI: 10.1137/16M1068177.

[Mél+18] J. Mélou, Y. Quéau, J.-D. Durou, F. Castan, and D. Cremers. “Variational Reflectance Estimation from Multi-View Images”. In: *Journal of Mathematical Imaging and Vision* 60.9 (2018), pp. 1527–1546. DOI: 10.1007/s10851-018-0809-x.

[Mes+19] M. Meshry, D. B. Goldman, S. Khamis, H. Hoppe, R. Pandey, N. Snavely, and R. Martin-Brualla. “Neural Rerendering in the Wild”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019. DOI: 10.1109/CVPR.2019.00704.

[Mil+20] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. “NeRF: Representing scenes as neural radiance fields for view synthesis”. In: *Proceedings of the European Conference on Computer Vision*. 2020. DOI: 10.1007/978-3-030-58452-8_24.

[MQ16] R. Mecca and Y. Quéau. “Unifying diffuse and specular reflections for the photometric stereo problem”. In: *IEEE Winter Conference on Applications of Computer Vision*. 2016. DOI: 10.1109/WACV.2016.7477643.

[MRC15] R. Mecca, E. Rodolà, and D. Cremers. “Realistic photometric stereo using partial differential irradiance equation ratios”. In: *Computers & Graphics* 51 (2015), pp. 8–16. DOI: 10.1016/j.cag.2015.05.020.

[Mun+22] J. Munkberg, J. Hasselgren, T. Shen, J. Gao, W. Chen, A. Evans, T. Mueller, and S. Fidler. “Extracting Triangular 3D Models, Materials, and Lighting From Images”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022. DOI: 10.1109/CVPR52688.2022.00810.

[Nam+18] G. Nam, J. H. Lee, D. Gutierrez, and M. H. Kim. “Practical SVBRDF acquisition of 3D objects with unstructured flash photography”. In: *ACM Transactions on Graphics* 37.6 (2018), 267:1–267:12. DOI: 10.1145/3272127.3275017.

[Neh+05] D. Nehab, S. Rusinkiewicz, J. Davis, and R. Ramamoorthi. “Efficiently combining positions and normals for precise 3D geometry”. In: *ACM Transactions on Graphics* 24.3 (2005), pp. 536–543. DOI: 10.1145/1073204.1073226.

[New+11] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohli, J. Shotton, S. Hodges, and A. Fitzgibbon. “KinectFusion: Real-time Dense Surface Mapping and Tracking”. In: *IEEE International Symposium on Mixed and Augmented Reality*. 2011. DOI: 10.1109/ISMAR.2011.6092378.

[Ngu+18] T. Nguyen-Phuoc, C. Li, S. Balaban, and Y. Yang. “RenderNet: A deep convolutional network for differentiable rendering from 3D shapes”. In: *Advances in Neural Information Processing Systems*. 2018. DOI: 10.5555/3327757.3327886.

[Nic+92] F. E. Nicodemus, J. C. Richmond, J. J. Hsia, I. W. Ginsberg, and T. Limperis. In: *Radiometry*. Jones and Bartlett Publishers, Inc., 1992. Chap. Geometrical Considerations and Nomenclature for Reflectance, pp. 94–145. DOI: 10.5555/136913.136929.

[Nie+13] M. Nießner, M. Zollhöfer, S. Izadi, and M. Stamminger. "Real-time 3D Reconstruction at Scale using Voxel Hashing". In: *ACM Transactions on Graphics*. 2013. DOI: 10.1145/2508363.2508374.

[Nie+20] M. Niemeyer, L. Mescheder, M. Oechsle, and A. Geiger. "Differentiable Volumetric Rendering: Learning Implicit 3D Representations without 3D Supervision". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020. DOI: 10.1109/CVPR42600.2020.00356.

[Nim+19] M. Nimier-David, D. Vicini, T. Zeltner, and W. Jakob. "Mitsuba 2: A Retargetable Forward and Inverse Renderer". In: *ACM Transactions on Graphics* 38.6 (2019), pp. 1–17. DOI: 10.1145/3355089.3356498.

[Nim+21] M. Nimier-David, Z. Dong, W. Jakob, and A. Kaplanyan. "Material and Lighting Reconstruction for Complex Indoor Scenes with Texture-space Differentiable Rendering". In: *Computer Graphics Forum*. 2021. DOI: 10.2312/sr.20211292.

[NJJ21] B. Nicolet, A. Jacobson, and W. Jakob. "Large Steps in Inverse Rendering of Geometry". In: *ACM Transactions on Graphics* 40.6 (2021), 248:1–248:13. DOI: 10.1145/3478513.3480501.

[NJR15] J. B. Nielsen, H. W. Jensen, and R. Ramamoorthi. "On optimal, minimal BRDF sampling for reflectance acquisition". In: *ACM Transactions on Graphics* 34.6 (2015), 186:1–186:11. DOI: 10.1145/2816795.2818085.

[ON14] G. Oxholm and K. Nishino. "Multiview Shape and Reflectance from Natural Illumination". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2014. DOI: 10.1109/CVPR.2014.277.

[ON94] M. Oren and S. K. Nayar. "Generalization of Lambert's Reflectance Model". In: *Proceedings of SIGGRAPH*. 1994, pp. 239–246. DOI: 10.1145/192161.192213.

[OPG21] M. Oechsle, S. Peng, and A. Geiger. "UNISURF: Unifying Neural Implicit Surfaces and Radiance Fields for Multi-View Reconstruction". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021. DOI: 10.1109/ICCV48922.2021.00554.

[Par+17] J. Park, S. N. Sinha, Y. Matsushita, Y. Tai, and I. S. Kweon. "Robust Multiview Photometric Stereo Using Planar Mesh Parameterization". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.8 (2017), pp. 1591–1604. DOI: 10.1109/TPAMI.2016.2608944.

[Pas+17] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. "Automatic Differentiation in PyTorch". In: *Advances in Neural Information Processing Systems Workshops*. 2017.

[Pas+18] D. Paschalidou, A. O. Ulusoy, C. Schmitt, L. van Gool, and A. Geiger. "RayNet: Learning Volumetric 3D Reconstruction with Ray Potentials". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018. DOI: 10.1109/CVPR.2018.00410.

[Pen+20] S. Peng, M. Niemeyer, L. Mescheder, M. Pollefeys, and A. Geiger. “Convolutional Occupancy Networks”. In: *Proceedings of the European Conference on Computer Vision*. 2020. DOI: 10.1007/978-3-030-58580-8_31.

[PF14] T. Papadhimitri and P. Favaro. “Uncalibrated Near-Light Photometric Stereo”. In: *British Machine Vision Conference*. 2014. DOI: 10.7892/boris.67317.

[Phi+19] J. Philip, M. Gharbi, T. Zhou, A. Efros, and G. Drettakis. “Multi-view Relighting Using a Geometry-Aware Network”. In: *ACM Transactions on Graphics* 38.4 (2019), 78:1–78:14. DOI: 10.1145/3306346.3323013.

[Phi+21] J. Philip, S. Morgenthaler, M. Gharbi, and G. Drettakis. “Free-Viewpoint Indoor Neural Relighting from Multi-View Stereo”. In: *ACM Transactions on Graphics* 40.5 (2021), 194:1–194:18. DOI: 10.1145/3469842.

[Pho75] B. T. Phong. “Illumination for Computer Generated Pictures”. In: *Communications of the ACM* 18.6 (1975), pp. 311–317. DOI: 10.1145/360825.360839.

[PHS20] J. J. Park, A. Holynski, and S. M. Seitz. “Seeing the World in a Bag of Chips”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020. DOI: 10.1109/CVPR42600.2020.00149.

[PJH16] M. Pharr, W. Jakob, and G. Humphreys. *Physically based rendering: From theory to implementation*. Morgan Kaufmann, 2016. DOI: 10.5555/3044800.

[PNS18] J. J. Park, R. A. Newcombe, and S. M. Seitz. “Surface Light Field Fusion”. In: *International Conference on 3D Vision*. 2018. DOI: 10.1109/3DV.2018.00013.

[Pol+08] M. Pollefeys, D. Nistér, J.-M. Frahm, A. Akbarzadeh, P. Mordohai, B. Clipp, C. Engels, D. Gallup, S. J. Kim, P. C. Merrell, C. Salmi, S. N. Sinha, B. Talton, L. Wang, Q. Yang, H. Stewénius, R. Yang, G. Welch, and H. Towles. “Detailed real-time urban 3d reconstruction from video”. In: *International Journal of Computer Vision* 78.2-3 (2008), pp. 143–167. DOI: 10.1007/s11263-007-0086-4.

[Pra+21] S. Prakash, T. Leimkühler, S. Rodriguez, and G. Drettakis. “Hybrid Image-Based Rendering for Free-View Synthesis”. In: *Proceedings of the ACM on Computer Graphics and Interactive Techniques* 4.1 (2021), pp. 1–20. DOI: 10.1145/3451260.

[Pra+22] S. Prakash, G. Rainer, A. Bousseau, and G. Drettakis. “Deep scene-scale material estimation from multi-view indoor captures”. In: *Computers & Graphics* 109 (2022), pp. 15–29. ISSN: 0097-8493. DOI: 10.1016/j.cag.2022.09.010.

[QLD15a] Y. Quéau, F. Lauze, and J. Durou. “A LÎ -TV Algorithm for Robust Perspective Photometric Stereo with Spatially-Varying Lightings”. In: *Scale Space and Variational Methods in Computer Vision*. 2015. DOI: 10.1007/978-3-319-18461-6_40.

[QLD15b] Y. Quéau, F. Lauze, and J. Durou. “Solving Uncalibrated Photometric Stereo Using Total Variation”. In: *Journal of Mathematical Imaging and Vision* 52.1 (2015), pp. 87–107. DOI: 10.1007/s10851-014-0512-5.

[QMD16] Y. Quéau, R. Mecca, and J. Durou. "Unbiased Photometric Stereo for Colored Surfaces: A Variational Approach". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2016. DOI: 10.1109/CVPR.2016.472.

[Qué+17a] Y. Quéau, J. Mélou, F. Castan, D. Cremers, and J. Durou. "A Variational Approach to Shape-from-Shading Under Natural Illumination". In: *Conference on Energy Minimization Methods in Computer Vision and Pattern Recognition*. 2017. DOI: 10.48550/arXiv.1709.10354.

[Qué+17b] Y. Quéau, J. Mélou, J. Durou, and D. Cremers. "Dense Multi-view 3D-reconstruction Without Dense Correspondences". In: *arXiv.org* (2017). DOI: 10.48550/arXiv.1704.00337.

[Qué+17c] Y. Quéau, T. Wu, F. Lauze, J. Durou, and D. Cremers. "A Non-convex Variational Approach to Photometric Stereo under Inaccurate Lighting". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2017. DOI: 10.1109/CVPR.2017.45.

[Qué+18] Y. Quéau, B. Durix, T. Wu, D. Cremers, F. Lauze, and J. Durou. "LED-Based Photometric Stereo: Modeling, Calibration and Numerical Solution". In: *Journal of Mathematical Imaging and Vision* 60.3 (2018), pp. 313–340. DOI: 10.1007/s10851-017-0761-1.

[QWC17] Y. Quéau, T. Wu, and D. Cremers. "Semi-calibrated Near-Light Photometric Stereo". In: *Scale Space and Variational Methods in Computer Vision*. 2017. DOI: 10.1007/978-3-319-58771-4_52.

[Rav+20] N. Ravi, J. Reizenstein, D. Novotny, T. Gordon, W.-Y. Lo, J. Johnson, and G. Gkioxari. "Accelerating 3D Deep Learning with PyTorch3D". In: *arXiv.org* (2020). DOI: 10.48550/arXiv.2007.08501.

[Ren+11] P. Ren, J. Wang, J. Snyder, X. Tong, and B. Guo. "Pocket Reflectometry". In: *ACM Transactions on Graphics* 30.4 (2011), 45:1–45:10. DOI: 10.1145/2010324.1964940.

[Riv+20] J. Riviere, P. Gotardo, D. Bradley, A. Ghosh, and T. Beeler. "Single-Shot High-Quality Facial Geometry and Skin Appearance Capture". In: *ACM Transactions on Graphics* 39.4 (2020), 81:1–81:12. DOI: 10.1145/3386569.3392464.

[RPG16] J. Rivière, P. Peers, and A. Ghosh. "Mobile Surface Reflectometry". In: *Computer Graphics Forum* 35.1 (2016), pp. 191–202. DOI: 10.1145/2614217.2630589.

[Rus98] S. Rusinkiewicz. "A New Change of Variables for Efficient BRDF Representation". In: *Eurographics Workshop on Rendering*. 1998. DOI: 10.1007/978-3-7091-6453-2_2.

[RV12] H. Roth and M. Vona. "Moving Volume KinectFusion". In: *British Machine Vision Conference*. 2012. DOI: 10.5244/C.26.112.

[Sch+13] C. Schwartz, R. Sarlette, M. Weinmann, and R. Klein. “DOME II: A Parallelized BTF Acquisition System”. In: *Eurographics Workshop on Material Appearance Modeling*. 2013. DOI: 10.5555/2600281.2600288.

[Sch+16] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm. “Pixelwise View Selection for Unstructured Multi-View Stereo”. In: *Proceedings of the European Conference on Computer Vision*. 2016. DOI: 10.1007/978-3-319-46487-9_31.

[Sch+20] C. Schmitt, S. Donné, G. Riegler, V. Koltun, and A. Geiger. “On Joint Estimation of Pose, Geometry and svBRDF from a Handheld Scanner”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020. DOI: 10.1109/CVPR42600.2020.00355.

[Sch+23] C. Schmitt, B. Antić, A. Neculai, J. H. Lee, and A. Geiger. “Towards Scalable Multi-View Reconstruction of Geometry and Materials”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023). DOI: 10.1109/TPAMI.2023.3314348.

[SD97] S. Seitz and C. Dyer. “Photorealistic scene reconstruction by voxel coloring”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1997. DOI: 10.1109/CVPR.1997.609462.

[SF16] J. L. Schönberger and J.-M. Frahm. “Structure-from-Motion Revisited”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2016. DOI: 10.1109/CVPR.2016.445.

[SFB03] D. Simakov, D. Frolova, and R. Basri. “Dense Shape Reconstruction of a Moving Object under Arbitrary, Unknown Lighting”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2003. DOI: 10.1109/ICCV.2003.1238628.

[Sha+21] M. Shafiei, S. Bi, Z. Li, A. Liaudanskas, R. O. Cayon, and R. Ramamoorthi. “Learning Neural Transmittance for Efficient Rendering of Reflectance Fields”. In: *British Machine Vision Conference*. 2021. DOI: 10.48550/arXiv.2110.13272.

[Shi+14] B. Shi, K. Inose, Y. Matsushita, P. Tan, S. Yeung, and K. Ikeuchi. “Photometric Stereo Using Internet Images”. In: *International Conference on 3D Vision*. 2014. DOI: 10.1109/3DV.2014.9.

[Shi+17] J. Shi, Y. Dong, H. Su, and S. X. Yu. “Learning Non-Lambertian Object Intrinsics Across ShapeNet Categories”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2017. DOI: 10.1109/CVPR.2017.619.

[Sit+19] V. Sitzmann, J. Thies, F. Heide, M. Nießner, G. Wetzstein, and M. Zollhöfer. “DeepVoxels: Learning Persistent 3D Feature Embeddings”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019. DOI: 10.1109/CVPR.2019.00254.

[SLS23] Z. Shi, X. Lin, and Y. Song. "An attention-embedded GAN for SVBRDF recovery from a single image". In: *Computational Visual Media* 9 (2023). DOI: 10.1007/s41095-022-0289-1.

[Sri+21] P. P. Srinivasan, B. Deng, X. Zhang, M. Tancik, B. Mildenhall, and J. T. Barron. "NeRV: Neural Reflectance and Visibility Fields for Relighting and View Synthesis". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021. DOI: 10.1109/CVPR46437.2021.00741.

[STL08] L. Shen, P. Tan, and S. Lin. "Intrinsic image decomposition with non-local texture cues". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2008. DOI: 10.1109/CVPR.2008.4587660.

[Str+19] J. Straub, T. Whelan, L. Ma, Y. Chen, E. Wijmans, S. Green, J. J. Engel, R. Mur-Artal, C. Ren, S. Verma, A. Clarkson, M. Yan, B. Budge, Y. Yan, X. Pan, J. Yon, Y. Zou, K. Leon, N. Carter, J. Briales, T. Gillingham, E. Mueggler, L. Pesqueira, M. Savva, D. Batra, H. M. Strasdat, R. D. Nardi, M. Goesele, S. Lovegrove, and R. Newcombe. "The Replica Dataset: A Digital Replica of Indoor Spaces". In: *arXiv.org* (2019). DOI: 10.48550/arXiv.1906.05797.

[Sun+23a] C. Sun, G. Cai, Z. Li, K. Yan, C. Zhang, C. Marshall, J.-B. Huang, S. Zhao, and Z. Dong. "Neural-PBIR Reconstruction of Shape, Material, and Illumination". In: *arXiv.org* (2023). DOI: 10.48550/arXiv.2304.13445.

[Sun+23b] J. Sun, Z. Zhang, T. Chu, G. Li, L. Zhao, and W. Xing. "Joint Optimization of Triangle Mesh, Material, and Light from Neural Fields with Neural Radiance Cache". In: *arXiv.org* (2023). DOI: 10.48550/arXiv.2305.16800.

[SYH13] L. Shen, C. Yeo, and B.-S. Hua. "Intrinsic Image Decomposition Using a Sparse Representation of Reflectance". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.12 (2013), pp. 2904–2915. DOI: 10.1109/TPAMI.2013.136.

[Tan+22] M. Tancik, V. Casser, X. Yan, S. Pradhan, B. Mildenhall, P. Srinivasan, J. T. Barron, and H. Kretzschmar. "Block-NeRF: Scalable Large Scene Neural View Synthesis". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022. DOI: 10.1109/CVPR52688.2022.00807.

[Tew+22] A. Tewari, J. Thies, B. Mildenhall, P. Srinivasan, E. Tretschk, W. Yifan, C. Lassner, V. Sitzmann, R. Martin-Brualla, S. Lombardi, T. Simon, C. Theobalt, M. Nießner, J. T. Barron, G. Wetzstein, M. Zollhöfer, and V. Golyanik. "Advances in Neural Rendering". In: *Computer Graphics Forum* 41.2 (2022), pp. 703–735. DOI: 10.1111/cgf.14507.

[Thi+20] J. Thies, M. Zollhöfer, C. Theobalt, M. Stamminger, and M. Nießner. "Image-guided Neural Object Rendering". In: *International Conference on Learning Representations*. 2020. DOI: 10.48550/arXiv.1811.10720.

[Tos+21] F. Tosi, Y. Liao, C. Schmitt, and A. Geiger. "SMD-Nets: Stereo Mixture Density Networks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021. DOI: 10.1109/CVPR46437.2021.00883.

[TS67] K. E. Torrance and E. M. Sparrow. “Theory for Off-Specular Reflection From Roughened Surfaces”. In: *Journal of the Optical Society of America* 57.9 (1967), pp. 1105–1114. DOI: 10.1364/JOSA.57.001105.

[Tun+13] B. Tunwattanapong, G. Fyffe, P. Graham, J. Busch, X. Yu, A. Ghosh, and P. E. Debevec. “Acquiring reflectance and shape from continuous spherical harmonic illumination”. In: *ACM Transactions on Graphics* 32.4 (2013), 109:1–109:12. DOI: 10.1145/2461912.2461944.

[UGB15] A. O. Ulusoy, A. Geiger, and M. J. Black. “Towards Probabilistic Volumetric Reconstruction using Ray Potentials”. In: *International Conference on 3D Vision*. 2015. DOI: 10.1109/3DV.2015.9.

[Vea98] E. Veach. “Robust Monte Carlo Methods for Light Transport Simulation”. PhD thesis. Stanford University, 1998. DOI: 10.5555/927297.

[Ver+22] D. Verbin, P. Hedman, B. Mildenhall, T. Zickler, J. T. Barron, and P. P. Srinivasan. “Ref-NeRF: Structured View-Dependent Appearance for Neural Radiance Fields”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022. DOI: 10.1109/CVPR52688.2022.00541.

[VSJ22] D. Vicini, S. Speierer, and W. Jakob. “Differentiable Signed Distance Function Rendering”. In: *ACM Transactions on Graphics* 41.4 (2022), 125:1–125:18. DOI: 10.1145/3528223.3530139.

[VTC05] G. Vogiatzis, P. H. S. Torr, and R. Cipolla. “Multi-View Stereo via Volumetric Graph-Cuts”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2005. DOI: 10.1109/CVPR.2005.238.

[Wan+20] J. Wang, V. Tantia, N. Ballas, and M. G. Rabbat. “SlowMo: Improving Communication-Efficient Distributed SGD with Slow Momentum”. In: *International Conference on Learning Representations*. 2020. DOI: 10.48550/arXiv.1910.00643.

[Wan+21a] P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang. “NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction”. In: *Advances in Neural Information Processing Systems*. 2021. DOI: 10.48550/arXiv.2106.10689.

[Wan+21b] Z. Wang, J. Philion, S. Fidler, and J. Kautz. “Learning Indoor Inverse Rendering with 3D Spatially-Varying Lighting”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021. DOI: 10.1109/ICCV48922.2021.01231.

[Wan+23] Z. Wang, T. Shen, J. Gao, S. Huang, J. Munkberg, J. Hasselgren, Z. Gojcic, W. Chen, and S. Fidler. “Neural Fields meet Explicit Geometric Representations for Inverse Rendering of Urban Scenes”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023. DOI: 10.48550/arXiv.2304.03266.

[Wei+21] Y. Wei, S. Liu, Y. Rao, W. Zhao, J. Lu, and J. Zhou. "NerfingMVS: Guided Optimization of Neural Radiance Fields for Indoor Multi-view Stereo". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021. DOI: 10.1109/ICCV48922.2021.00556.

[Wen+22] T. Wen, B. Wang, L. Zhang, J. Guo, and N. Holzschuch. "SVBRDF Recovery from a Single Image with Highlights Using a Pre-trained Generative Adversarial Network". In: *Computer Graphics Forum* 41.6 (2022), pp. 110–123. DOI: 10.1111/cgf.14514.

[Whe+12] T. Whelan, M. Kaess, M. F. andH. Johannsson, J. Leonard, and J. McDonald. "Kintinuous: Spatially Extended KinectFusion". In: *RSS Workshop on RGB-D: Advanced Reasoning with Depth Cameras*. 2012.

[Whe+15] T. Whelan, S. Leutenegger, R. F. Salas-Moreno, B. Glocker, and A. J. Davison. "ElasticFusion: Dense SLAM Without A Pose Graph". In: *Proceedings of Robotics: Science and Systems*. 2015. DOI: 10.15607/RSS.2015.XI.001.

[Whe+16] T. Whelan, R. Salas-Moreno, B. Glocker, A. Davison, and S. Leutenegger. "ElasticFusion: Real-time dense SLAM and light source estimation". In: *The International Journal of Robotics Research* 35 (2016), pp. 1697–1716. DOI: 10.1177/0278364916669237.

[Wil+20] O. Wiles, G. Gkioxari, R. Szeliski, and J. Johnson. "SynSin: End-to-End View Synthesis From a Single Image". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020. DOI: 10.1109/CVPR42600.2020.00749.

[WJP21] F. R. Wasee, A. Joy, and C. Poullis. "Predicting Surface Reflectance Properties of Outdoor Scenes Under Unknown Natural Illumination". In: *arXiv.org* (2021). DOI: 10.48550/arXiv.2105.06820.

[WO16] J. Wang and E. Olson. "AprilTag 2: Efficient and robust fiducial detection". In: *IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2016. DOI: 10.1109/IROS.2016.7759617.

[Woo+00] D. N. Wood, D. I. Azuma, K. Aldinger, B. Curless, T. Duchamp, D. Salesin, and W. Stuetzle. "Surface light fields for 3D photography". In: *ACM Transactions on Graphics*. 2000, pp. 287–296. DOI: 10.1145/344779.344925.

[Woo80] R. J. Woodham. "Photometric method for determining surface orientation from multiple images". In: *Optical Engineering* 19.1 (1980), p. 191139. DOI: 10.5555/93871.93888.

[Wu+14] C. Wu, M. Zollhöfer, M. Nießner, M. Stamminger, S. Izadi, and C. Theobalt. "Real-time Shading-based Refinement for Consumer Depth Cameras". In: *ACM Transactions on Graphics*. Vol. 33. 6. 2014. DOI: 10.1145/2661229.2661232.

[Wu+22] X. Wu, J. Xu, Z. Zhu, H. Bao, Q. Huang, J. Tompkin, and W. Xu. "Scalable Neural Indoor Scene Rendering". In: *ACM Transactions on Graphics* 41.4 (2022), 98:1–98:16. DOI: 10.1145/3528223.3530153.

[Wu+23a] H. Wu, Z. Hu, L. Li, Y. Zhang, C. Fan, and X. Yu. “NeFII: Inverse Rendering for Reflectance Decomposition with Near-Field Indirect Illumination”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023. DOI: 10.48550/arXiv.2303.16617.

[Wu+23b] L. Wu, R. Zhu, M. B. Yaldiz, Y. Zhu, H. Cai, J. Matai, F. Porikli, T.-M. Li, M. Chandraker, and R. Ramamoorthi. “Factorized Inverse Path Tracing for Efficient and Accurate Material-Lighting Estimation”. In: *arXiv.org* (2023). DOI: 10.48550/arXiv.2304.05669.

[WWR22] F. Wimbauer, S. Wu, and C. Rupprecht. “De-rendering 3D Objects in the Wild”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022. DOI: 10.1109/CVPR52688.2022.01794.

[WWZ16] H. Wu, Z. Wang, and K. Zhou. “Simultaneous Localization and Appearance Estimation with a Consumer RGB-D Camera”. In: *IEEE Transactions on Visualization and Computer Graphics* 22.8 (2016), pp. 2012–2023. DOI: 10.1109/TVCG.2015.2498617.

[WYT16] Z. Wu, S. Yeung, and P. Tan. “Towards Building an RGBD-M Scanner”. In: *arXiv.org* (2016). DOI: 10.48550/arXiv.1603.03875.

[WZ15] H. Wu and K. Zhou. “AppFusion: Interactive Appearance Acquisition Using a Kinect Sensor”. In: *Computer Graphics Forum* 34.6 (2015), pp. 289–298. DOI: 10.1111/cgf.12600.

[XDW15] W. Xie, C. Dai, and C. C. L. Wang. “Photometric stereo with near point lighting: A solution by mesh deformation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2015. DOI: 10.1109/CVPR.2015.7299089.

[Xia+16] R. Xia, Y. Dong, P. Peers, and X. Tong. “Recovering shape and spatially-varying surface reflectance under unknown illumination”. In: *ACM Transactions on Graphics* 35.6 (2016), 187:1–187:12. DOI: 10.1145/2980179.2980248.

[Xin+23] X. Xing, K. Groh, S. Karaoglu, and T. Gevers. “Intrinsic Appearance Decomposition Using Point Cloud Representation”. In: *arXiv.org* (2023). DOI: 10.48550/arXiv.2307.10924.

[Xu+16] Z. Xu, J. B. Nielsen, J. Yu, H. W. Jensen, and R. Ramamoorthi. “Minimal BRDF sampling for two-shot near-field reflectance acquisition”. In: *ACM Transactions on Graphics* 35.6 (2016), 188:1–188:12. DOI: 10.1145/2980179.2982396.

[Xu+18] Z. Xu, K. Sunkavalli, S. Hadap, and R. Ramamoorthi. “Deep Image-Based Relighting from Optimal Sparse Samples”. In: *ACM Transactions on Graphics* 37.4 (2018), pp. 1–13. DOI: 10.1145/3197517.3201313.

[Xu+19] Z. Xu, S. Bi, K. Sunkavalli, S. Hadap, H. Su, and R. Ramamoorthi. “Deep View Synthesis from Sparse Photometric Images”. In: *ACM Transactions on Graphics* 38.4 (2019), 76:1–76:13. DOI: 10.1145/3306346.3323007.

[Yan+20] S. Yang, B. Li, Y.-P. Cao, H. Fu, Y.-K. Lai, L. Kobbelt, and S.-M. Hu. "Noise-Resilient Reconstruction of Panoramas and 3D Scenes Using Robot-Mounted Unsynchronized Commodity RGB-D Cameras". In: *ACM Transactions on Graphics* 39.5 (2020), pp. 1–15. DOI: 10.1145/3389412.

[Yan+22] W. Yang, G. Chen, C. Chen, Z. Chen, and K.-Y. K. Wong. "PS-NeRF: Neural Inverse Rendering for Multi-view Photometric Stereo". In: *Proceedings of the European Conference on Computer Vision*. 2022. DOI: 10.1007/978-3-031-19769-7_16.

[Yao+22] Y. Yao, J. Zhang, J. Liu, Y. Qu, T. Fang, D. McKinnon, Y. Tsin, and L. Quan. "NeILF: Neural Incident Light Field for Physically-based Material Estimation". In: *Proceedings of the European Conference on Computer Vision*. 2022. DOI: 10.1007/978-3-031-19821-2_40.

[Yar+20] L. Yariv, Y. Kasten, D. Moran, M. Galun, M. Atzmon, B. Ronen, and Y. Lipman. "Multiview Neural Surface Reconstruction by Disentangling Geometry and Appearance". In: *Advances in Neural Information Processing Systems*. 2020. DOI: 10.5555/3495724.3495934.

[Yar+21] L. Yariv, J. Gu, Y. Kasten, and Y. Lipman. "Volume rendering of neural implicit surfaces". In: *Advances in Neural Information Processing Systems*. 2021. DOI: 10.48550/arXiv.2106.12052.

[Yif+19] W. Yifan, F. Serena, S. Wu, C. Öztireli, and O. Sorkine-Hornung. "Differentiable surface splatting for point-based geometry processing". In: *ACM Transactions on Graphics* 38 (2019), pp. 1–14. DOI: 10.1145/3355089.3356513.

[YN23] K. Yoshiyama and T. Narihira. "NDJIR: Neural Direct and Joint Inverse Rendering for Geometry, Lights, and Materials of Real Object". In: *arXiv.org* (2023). DOI: 10.48550/arXiv.2302.00675.

[Yu+99] Y. Yu, P. E. Debevec, J. Malik, and T. Hawkins. "Inverse Global Illumination: Recovering Reflectance Models of Real Scenes from Photographs". In: *ACM Transactions on Graphics*. 1999, pp. 215–224. DOI: 10.1145/311535.311559.

[YY11] Y. Yoshiyasu and N. Yamazaki. "Topology-adaptive multi-view photometric stereo". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2011. DOI: 10.1109/CVPR.2011.5995576.

[YYH20] Z. Yang, S. Yan, and Q. Huang. "Extreme Relative Pose Network Under Hybrid Representations". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020. DOI: 10.1109/CVPR42600.2020.00253.

[Zen+13] M. Zeng, F. Zhao, J. Zheng, and X. Liu. "Octree-Based Fusion for Realtime 3D Reconstruction". In: *Graphical Models* 75.3 (2013), pp. 126–136. DOI: 10.1016/j.gmod.2012.09.002.

[Zen+17] A. Zeng, S. Song, M. Nießner, M. Fisher, J. Xiao, and T. Funkhouser. "3DMatch: Learning Local Geometric Descriptors from RGB-D Reconstructions". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2017. DOI: 10.1109/CVPR.2017.29.

[Zha+20] K. Zhang, G. Riegler, N. Snavely, and V. Koltun. "NeRF++: Analyzing and improving neural radiance fields". In: *arXiv.org* (2020). DOI: 10.48550/arXiv.2010.07492.

[Zha+21a] J. Y. Zhang, G. Yang, S. Tulsiani, and D. Ramanan. "NeRS: Neural Reflectance Surfaces for Sparse-view 3D Reconstruction in the Wild". In: *Advances in Neural Information Processing Systems*. 2021. DOI: 10.48550/arXiv.2110.07604.

[Zha+21b] K. Zhang, F. Luan, Q. Wang, K. Bala, and N. Snavely. "PhySG: Inverse Rendering with Spherical Gaussians for Physics-based Material Editing and Relighting". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021. DOI: 10.1109/CVPR46437.2021.00541.

[Zha+21c] X. Zhang, P. P. Srinivasan, B. Deng, P. Debevec, W. T. Freeman, and J. T. Barron. "NeRFactor: Neural Factorization of Shape and Reflectance under an Unknown Illumination". In: *ACM Transactions on Graphics* 40.6 (2021), pp. 1–18. DOI: 10.1145/3478513.3480496.

[Zha+22] Y. Zhang, J. Sun, X. He, H. Fu, R. Jia, and X. Zhou. "Modeling Indirect Illumination for Inverse Rendering". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022. DOI: 10.1109/CVPR52688.2022.01809.

[Zha+23a] J. Zhang, Y. Yao, S. Li, J. Liu, T. Fang, D. McKinnon, Y. Tsin, and L. Quan. "NeILF++: Inter-Reflectable Light Fields for Geometry and Material Estimation". In: *arXiv.org* (2023). DOI: 10.48550/arXiv.2303.17147.

[Zha+23b] L. Zhang, F. Gao, L. Wang, M. Yu, J. Cheng, and J. Zhang. "Deep SVBRDF Estimation from Single Image under Learned Planar Lighting". In: *ACM Transactions on Graphics*. 2023. DOI: 10.1145/3588432.3591559.

[Zha+23c] Y. Zhang, T. Xu, J. Yu, Y. Ye, J. Wang, Y. Jing, J. Yu, and W. Yang. "NeMF: Inverse Volume Rendering with Neural Microflake Field". In: *arXiv.org* (2023). DOI: 10.48550/arXiv.2304.00782.

[Zha+23d] Z. Zhang, R. Peng, Y. Hu, and R. Wang. "GeoMVSNet: Learning Multi-View Stereo With Geometry Perception". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

[Zha+23e] D. Zhao, D. Lichy, P.-N. Perrin, J.-M. Frahm, and S. Sengupta. "MVPSNet: Fast Generalizable Multi-view Photometric Stereo". In: *arXiv.org* (2023). DOI: 10.48550/arXiv.2305.11167.

[Zha+99] R. Zhang, P. Tsai, J. E. Cryer, and M. Shah. "Shape from Shading: A Survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 21.8 (1999), pp. 690–706. DOI: 10.1109/34.784284.

[Zha99] Z. Zhang. “Flexible Camera Calibration by Viewing a Plane from Unknown Orientations”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1999. DOI: 10.1109/ICCV.1999.791289.

[Zho+16] Z. Zhou, G. Chen, Y. Dong, D. Wipf, Y. Yu, J. Snyder, and X. Tong. “Sparse-as-possible SVBRDF Acquisition”. In: *ACM Transactions on Graphics* 35.6 (2016), 189:1–189:12. DOI: 10.1145/2980179.2980247.

[Zho+21] B. Zhou, W. Ma, Q. Li, N. El-Sheimy, Q. Mao, Y. Li, F. Gu, L. Huang, and J. Zhu. “Crowdsourcing-based indoor mapping using smartphones: A survey”. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 177 (May 2021), pp. 131–146. DOI: 10.1016/j.isprsjprs.2021.05.006.

[Zhu+22a] J. Zhu, F. Luan, Y. Huo, Z. Lin, Z. Zhong, D. Xi, R. Wang, H. Bao, J. Zheng, and R. Tang. “Learning-Based Inverse Rendering of Complex Indoor Scenes with Differentiable Monte Carlo Raytracing”. In: *ACM Transactions on Graphics*. 2022. DOI: 10.1145/3550469.3555407.

[Zhu+22b] R. Zhu, Z. Li, J. Matai, F. Porikli, and M. Chandraker. “IRISformer: Dense Vision Transformers for Single-Image Inverse Rendering in Indoor Scenes”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022. DOI: 10.1109/CVPR52688.2022.00284.

[Zhu+23a] J. Zhu, Y. Huo, Q. Ye, F. Luan, J. Li, D. Xi, L. Wang, R. Tang, W. Hua, H. Bao, and R. Wang. “I^2-SDF: Intrinsic Indoor Scene Reconstruction and Editing via Raytracing in Neural SDFs”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023. DOI: 10.48550/arXiv.2303.07634.

[Zhu+23b] S. Zhu, S. Saito, A. Bozic, C. Aliaga, T. Darrell, and C. Lassner. “Neural Relighting with Subsurface Scattering by Learning the Radiance Transfer Gradient”. In: *arXiv.org* (2023). DOI: 10.48550/arXiv.2306.09322.

[ZK13] Q.-Y. Zhou and V. Koltun. “Dense Scene Reconstruction with Points of Interest”. In: *ACM Transactions on Graphics* 32.4 (2013), pp. 1–8. DOI: 10.1145/2461912.2461919.

[ZK14] Q. Zhou and V. Koltun. “Color map optimization for 3D reconstruction with consumer depth cameras”. In: *ACM Transactions on Graphics*. Vol. 33. 4. 2014. DOI: 10.1145/2601097.2601134.

[ZK21] X. Zhou and N. K. Kalantari. “Adversarial Single-Image SVBRDF Estimation with Hybrid Training”. In: *Computer Graphics Forum* 40.2 (2021), pp. 315–325. DOI: 10.1111/cgf.142635.

[ZK22] X. Zhou and N. K. Kalantari. “Look-Ahead Training with Learned Reflectance Loss for Single-Image SVBRDF Estimation”. In: *ACM Transactions on Graphics* 41.6 (2022). DOI: 10.1145/3550454.3555495.

[ZMK13] Q.-Y. Zhou, S. Miller, and V. Koltun. “Elastic Fragments for Dense Scene Reconstruction”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2013. DOI: 10.1109/ICCV.2013.65.

[Zol+15] M. Zollhöfer, A. Dai, M. Innmann, C. Wu, M. Stamminger, C. Theobalt, and M. Nießner. “Shading-based refinement on volumetric signed distance functions”. In: *ACM Transactions on Graphics* 34.4 (2015), 96:1–96:14. DOI: 10.1145/2766887.

[ZT10] Z. Zhou and P. Tan. “Ring-Light Photometric Stereo”. In: *Proceedings of the European Conference on Computer Vision*. 2010. DOI: 10.1007/978-3-642-15552-9_20.

[Zuo+17] X. Zuo, S. Wang, J. Zheng, and R. Yang. “Detailed Surface Geometry and Albedo Recovery from RGB-D Video under Natural Illumination”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2017. DOI: 10.1109/ICCV.2017.340.

[ZWT13] Z. Zhou, Z. Wu, and P. Tan. “Multi-view Photometric Stereo with Spatially Varying Isotropic Materials”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2013. DOI: 10.1109/CVPR.2013.195.

[ZZL23] Y. Zhang, J. Zhu, and L. Lin. “Multi-View Stereo Representation Revist: Region-Aware MVSNet”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023. DOI: 10.48550/arXiv.2304.13614.