TESIS DE LA UNIVERSIDAD DE ZARAGOZA
2026
17

Víctor Martínez Batlle

Real-Scale Dense Monocular SLAM Exploiting Illumination Decline

Director/es
Tardos Solano, Juan Domingo

ISSN 2254-7606

Prensas de la Universidad
Universidad Zaragoza

Universidad de Zaragoza
Servicio de Publicaciones

ISSN 2254-7606

Universidad Zaragoza
1542

Tesis Doctoral

REAL-SCALE DENSE MONOCULAR SLAM EXPLOITING ILLUMINATION DECLINE

Autor

Víctor Martínez Batlle

Director/es

Tardos Solano, Juan Domingo

UNIVERSIDAD DE ZARAGOZA
Escuela de Doctorado

2025

Universidad Zaragoza

1474

Tesis Doctoral

REAL-SCALE DENSE MONOCULAR SLAM EXPLOITING ILLUMINATION DECLINE

Autor

Víctor Martínez Batlle

Director

Juan Domingo Tardós Solano

Programa de Doctorado en Ingeniería de Sistemas e Informática
Escuela de Doctorado

2025

Acknowledgements

I would like to thank Prof. Juan Domingo Tardós and Prof. José María Martínez Montiel for their help and support over the past four years as I've forged my path as a researcher. Mingo, thank you for sharing with me your incredible intuition, advice and guidance. You are an unbeatable supervisor. Monti, thank you for passing on your passion for research, always striving for the next step. To both of you: THANK YOU.

I would also like to thank Lina Paz-Perez and Prof. Pascal Fua for the opportunity to collaborate with their outstanding teams. Lina, my time at Apple has truly been the icing on the cake of my PhD journey. Pascal, working with you in Zaragoza and at EPFL has always been fun—but above all, it has been an honor and a pleasure.

These years would not have been the same without all the people who have accompanied me across Zaragoza, Japan, California, Canada, Switzerland... Gracias a mi familia, tíos, primas, mamá, papá, Clara, gracias por aguantar mi ironía, en mis ratos buenos y malos. Montse, Patri, gracias por una vida juntos y por cada tarde de merienda y juegos. Os quiero. Alex, Alicia, Pilar, Carol, Jorge, cada vez que nos vemos acabo con una sonrisa. Mucha, mucha suerte en vuestro propio camino. Diego, qué te voy a contar, me tocó la maldita lotería al conocerte. Edurne, David, Jorge, Néstor, Dario, Pablo, Juan Raúl, Martincho y Fer, prometedme más viajes juntos. El tiempo con vosotros nunca me parece suficiente. Juanjo, Julia, Sergio, ha sido el mejor verano de estos años, ojalá repetirlo. Javier R, Morlana, Tirado, Lorenzo, Tomás, Samu, Nico, Bruno, Raúl, Xavi, los del fondo del pasillo, sois los mejores, gracias por amenizar cada día de trabajo. Mi fantástica gente de robótica y gráficos, junior y sénior, todos los nombres no caben aquí, gracias por todo el buen rollo. My people in Switzerland: Miguel, Elías, Albert, Andrés, Carla, Vanesa, Solène, Corentin, Aoxiang. My people in the US: Chema, Joe, Jai, Cal, José. You made me feel right at home. Last but certainly not least, I sincerely thank all my colleagues at EndoMapper and at the Hospital Clínico de Zaragoza.

A todos vosotros, gracias de corazón.

Resumen

La localización y mapeo visual simultáneos (Visual SLAM o VSLAM) es un campo consolidado para la reconstrucción de escenas en interiores y exteriores, pero su aplicación a la endoscopia médica sigue estando en gran medida inexplorada. Los vídeos de endoscopia suelen presentar escasa textura y paralaje. Además, la única fuente de luz en la escena está rígidamente unida a la cámara y se mueve generando continuas y complejas variaciones de iluminación. Estos cambios de brillo se han tratado tradicionalmente como una molestia y, en su mayoría, se han ignorado. En esta tesis, convertimos la limitación en oportunidad: somos los primeros en aprovechar los efectos del decaimiento de la luz para obtener reconstrucciones 3D densas y precisas a partir de vídeo monocular en procedimientos médicos reales de colonoscopia.

Introducimos un modelo fotométrico específicamente diseñado para vídeo endoscópico, que captura la relación entre las fuentes de luz, la escena, la cámara y las imágenes capturadas. Esta contribución permite, por primera vez, la reconstrucción in-vivo de la superficie 3D del colon humano con un error medio inferior a los 3 mm.

Nuestra novedosa red de estimación de profundidad es la primera red autosupervisada monovista; y predice, además, las normales y el albedo de cada punto a partir de una única imagen. Esto permitirá a un sistema VSLAM generar mapas de densos y precisos con errores promedio de tan solo un 7–8 %, sin anotaciones de entrenamiento.

A continuación, nos centramos en métodos multivista para recuperar, por primera vez, una superficie implícita, densa y estanca del colon humano a partir de vídeo. Esta es lograda mediante la optimización de una función de distancia con signo (SDF) que representa la forma 3D inherentemente estanca del colon.

Los sistemas monoculares suelen estar limitados por la ambigüedad de escala. Abordamos este desafío final con nuestro método: el primero en aprovechar la distancia conocida entre la cámara y las fuentes de luz para estimar la escala con precisión milimétrica. Esto convierte efectivamente cualquier endoscopio monocular en un sensor 3D métrico. El método alcanza precisión submilimétrica en simulación y tiene un rendimiento al medir pólipos comparable al de endoscopistas expertos.

En resumen, esta tesis sienta las bases para aprovechar la iluminación en la endoscopia clínica estándar. Como resultado, podemos obtener reconstrucciones 3D y SLAM monocular métrico y denso; no compensando las variaciones de brillo, sino aprovechándolas como una valiosa fuente de información.

Abstract

Visual SLAM (VSLAM) is a mature field for indoor and outdoor scene reconstruction, but its application to medical endoscopy remains yet largely underexplored. Endoscopic videos typically exhibit limited texture and parallax, and, critically, the only light source in the scene is rigidly attached to the moving camera, leading to continuous and complex lighting variations. These illumination changes have traditionally been treated as a nuisance and largely ignored. In this thesis, we turn this limitation into an opportunity: we are the first to exploit the effects of near-field illumination decline to enable accurate and dense 3D reconstruction from monocular colonoscopy video in real medical procedures.

We introduce a photometric model specifically tailored to endoscopic video, capturing the relationship between the light sources, the scene, the camera and the images. This core contribution enables, for the first time, the in-vivo reconstruction of the human colon 3D surface with a mean error below 3 mm.

Our novel single-view, self-supervised depth estimation network predicts depth, surface normals, and spatially varying albedo from a single image. This method would empower any VSLAM system to produce dense, high-accuracy depth maps, achieving average errors of just 7–8%, all without requiring annotated training data.

We then bridge the gap from single-view to multi-view dense reconstruction by recovering, for the first time, a watertight implicit neural surface of the human colon from endoscopic video. This is accomplished by optimizing a signed distance function (SDF) that implicitly represents the inherently watertight 3D shape of the colon.

Monocular systems are typically limited by scale ambiguity. We address this final challenge with our method: the first to leverage the known baseline between the camera and light sources to estimate scale with millimeter-level accuracy. This effectively transforms any monocular endoscope into a fully metric 3D sensing device. The method achieves sub-millimeter precision in simulation and performs comparably to experienced endoscopists when measuring polyps.

In summary, this thesis lays the groundwork for exploiting illumination in standard clinical endoscopy. As a result, it enables dense, metric, monocular SLAM and 3D reconstruction—not by compensating for brightness variations, but by embracing them as a powerful source of information.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Visual Simultaneous Localization and Mapping (Visual SLAM or VSLAM) is a widely studied problem in robotics and computer vision. It involves the real-time estimation of a camera's trajectory while simultaneously constructing a 3D map of the surrounding environment, using only visual input. Among the various approaches developed, keypoint-based methods (Davison et al., 2007; Klein and Murray, 2007; Mur-Artal et al., 2015; Campos et al., 2021) have played a central role, delivering robust performance across a wide range of applications.

Among these applications, a new and exciting direction emerges from the goal of bringing 3D mapping capabilities into the human body (Schmidt et al., 2024). These maps can support a variety of clinical tasks such as enhanced surgical navigation, augmented reality overlays, standardized quality assessment during exploratory procedures, tumor localization, targeted biopsies, or highly accurate drug delivery during interventions.

Optical endoscopy is the gold standard for minimally invasive diagnostics and interventions, providing direct visualization of internal anatomy while enabling therapeutic procedures in gastroenterology (East et al., 2016) and pulmonology (Criner et al., 2020). Clinical endoscopes are designed as slender, flexible tubes that must navigate the narrow anatomy of the human body. Their maneuverability is inherently constrained, resulting in small variations in viewpoint and minimal parallax between frames—challenging conditions for SLAM and structure-from-motion (SfM) techniques. At the distal end of the scope, spatial limitations impose a strict design budget, typically allowing only a single monocular camera. Alongside it, auxiliary components are integrated: a water jet for cleansing the mucosa, a working channel for surgical tools, and—critically for our purposes—two or more miniature light sources. These are the sole sources of illumination and move with the camera, enabling us to exploit near-field lighting to enhance VSLAM in endoscopy.

Despite their success in many domains, visual SLAM systems face significant chal-

lenges when applied within the human body. These include limited texture, insufficient parallax, dynamic tissue deformation, and varying illumination conditions. Illumination, in particular, is often overlooked, with many methods assuming brightness constancy such as RNNSLAM (Ma et al., 2021), EndoGSLAM (Wang et al., 2024b) and ENeRF-SLAM (Shan et al., 2024) or relying on brightness-invariant tracking as in DefSLAM (Lamarca et al., 2020), EndoSLAM (Ozyoruk et al., 2021) and NR-SLAM (Gómez-Rodríguez et al., 2024). Moreover, the limitation to monocular cameras results in scale ambiguity due to the absence of direct depth measurements. Consequently, the resulting maps are only accurate up to scale, which restricts their utility in applications requiring detailed and metric reconstructions, such as lesion measurements, preoperative planning, or intraoperative guidance.

This thesis focuses on *Visual SLAM for endoscopy*, targeting environments characterized by *near-field artificial illumination*, where the light source is close to the surface. The light is also co-located with the camera and moves in tandem with it. These conditions are not unique to medical endoscopy. They also arise in applications such as nighttime driving, deep-sea exploration, and underground robotics. All of these scenarios present a unique opportunity to exploit lighting cues for metric 3D reconstruction. Specifically, we use the term *illumination decline* to refer to the set of light transport phenomena that can be used to obtain additional geometric information from the scene.

Contributions

1.1 Photometric model

In general, we consider a setup involving a monocular endoscope equipped with two or more *light sources* positioned near the *camera* at the distal end (tip) of the endoscope as shown in Figure 1.1. The light sources are the only source of illumination in the environment, and they are typically modeled as point or spotlight sources. The camera has a wide field of view (FOV), achieved using a fish-eye lens.

Our photometric model for endoscopic imaging exploits a key effect of near-light illumination, the *inverse-square law* of illumination decline:

$$\mathcal{I} \propto \frac{1}{d^2} \tag{1.1}$$

where the intensity of light perceived by the camera $\mathcal{I}$ decreases with the square of the distance d from the source. This effect is crucial for estimating depth from the observed brightness of a surface. See Figure 1.1 for an overview of the model.

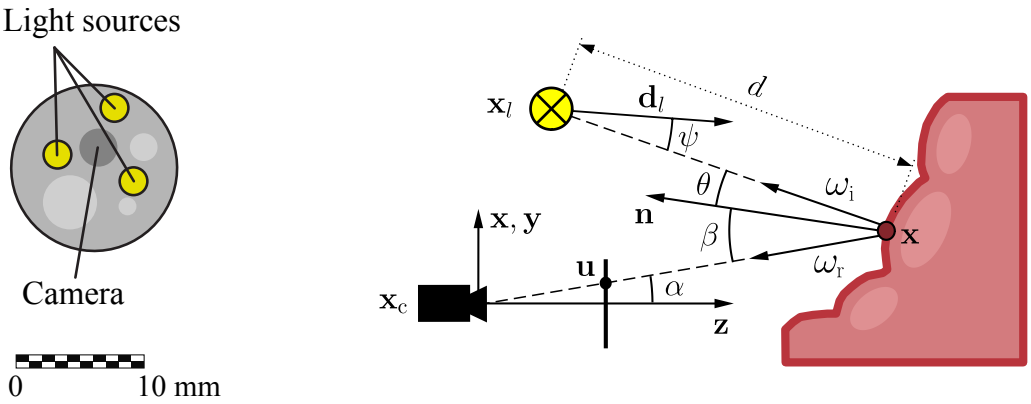


Figure 1.1: **Left:** Schematic representation of the endoscope tip. **Right:** Overview of the proposed photometric model. See chapter 2 for a detailed discussion.

	Photometric model	**Method**	**Depth**	**Normal**	**Albedo**	**Multi view**	**Metric scale**
Baseline		Non-linear optimization	✓	✓	-	-	-
LightDepth		Deep neural networks	✓	✓	✓	-	-
LightNeuS		Radiance fields	✓	✓	✓	✓	-
EndoMetric		Non-linear optimization	✓	✓	✓	✓	✓

Table 1.1: Summary of the different 3D reconstruction methods proposed

1.2 Reconstruction methods

This thesis proposes four different methods for 3D reconstruction in endoscopy, summarized in Table 1.1. Note that the first three methods are based on a simplified photometric model that assumes a single *virtual light source* representing the combined effects of the real light sources.

Our methods also differ primarily in their assumptions about surface reflectance. The baseline method and EndoMetric adopt a diffuse, or *Lambertian*, surface model, where light is reflected uniformly in all directions. In contrast, LightDepth and LightNeuS account for non-Lambertian effects, such as specular reflections. All methods estimate a per-point surface color, or *albedo*, except for the baseline, which assumes a constant uniform reflectance.

1.2.1 Baseline photometric method for 3D reconstruction in colonoscopy with millimeter accuracy

In chapter 2, we exploit the controlled lighting in colonoscopy to achieve the first in-vivo 3D reconstruction of the human colon using a calibrated monocular endoscope. Our method works in a real medical environment, providing both a suitable in-place calibration procedure and a depth estimation technique adapted to the colon's tubular geometry. We validate our method on simulated colonoscopies, obtaining a mean error of 7% on depth estimation, which is below 3 mm on average. Our qualitative results on the EndoMapper dataset show that the method is able to correctly estimate the colon shape in real human colonoscopies, paving the ground for true-scale monocular SLAM in endoscopy.

1.2.2 LightDepth: A single-view self-supervised depth estimation method as accurate as fully supervised approaches

Our baseline method in chapter 2 assumed constant albedo and was based on an optimization that might be too slow for real-time applications. To overcome these limitations, we explore the use of neural networks to estimate depth and albedo from a single image. In chapter 3, we introduce the first single-view, self-supervised depth estimation method for endoscopy. Our approach achieves performance comparable to fully supervised methods, yet requires no ground-truth depth data. The key contribution of this work lies in leveraging illumination decline and photometric calibration of the endoscope to derive a strong supervisory signal. Notably, our method can be trained on large colonoscopy datasets without any ground-truth labels and can even be refined online at test time.

1.2.3 LightNeuS: A method to obtain watertight neural implicit surface reconstructions

The previous method cannot leverage multi-view information to improve 3D reconstruction. In chapter 4, we introduce a neural multi-view approach for surface reconstruction from sequences of images acquired with monocular endoscopes. It is based on a new key insight: endoluminal cavities are watertight, a property naturally enforced by modeling them in terms of a signed distance function. To exploit this insight, we build on NeuS (Wang et al., 2021b), a neural implicit surface reconstruction technique with an outstanding capability to learn appearance and an SDF surface model from multiple views, but currently limited to scenes with static illumination. To remove this limitation and exploit the relation between pixel brightness and depth, we modify the

NeuS architecture to explicitly account for it and introduce a calibrated photometric model of the endoscope's camera and light source.

Our method is the first one to produce watertight reconstructions of whole colon sections. We demonstrate excellent accuracy on phantom imagery. Remarkably, the watertight prior combined with illumination decline, allows to complete the reconstruction of unseen portions of the surface with acceptable accuracy, paving the way to automatic quality assessment of cancer screening explorations, measuring the global percentage of observed mucosa.

1.2.4 EndoMetric: A method to recover metric scale from a monocular reconstruction with a near-light model

All previous methods yield only up-to-scale reconstructions. With a moving monocular camera, the absolute scale of the environment is inherently unobservable, resulting in 3D reconstructions and camera trajectories that are determined only up to an unknown scale factor. For the first time, we propose in chapter 5 a method to estimate the real metric scale of a 3D reconstruction from standard monocular endoscopic images, under unknown varying albedo, without relying on application-specific learned priors. Our fully model-based approach leverages the near-light sources embedded in endoscopes, positioned at a small but nonzero baseline from the camera, in combination with the inverse-square law of light attenuation, to accurately recover the metric scale from scratch. This enables the transformation of any endoscope into a metric device, which is crucial for applications such as measuring polyps, stenosis, or assessing the extent of diseased tissue.

1.3 Results

1.3.1 Publications

- V. M. Batlle, J. M. M. Montiel, and J. D. Tardós. Photometric single-view dense 3D reconstruction in endoscopy. In *IEEE Int. Conf. on Intelligent Robots and Systems (IROS)*, pages 4904–4910, 2022.

- P. Azagra, C. Sostres, Á. Ferrandez, L. Riazuelo, C. Tomasini, O. L. Barbed, J. Morlana, D. Recasens, V. M. Batlle, J. J. Gómez-Rodríguez, R. Elvira, J. López, C. Oriol, J. Civera, J. D. Tardós, A. C. Murillo, A. Lanas, and J. M. M. Montiel. Endomapper dataset of complete calibrated endoscopy procedures. *Scientific Data*, 10(1):671, 2023.

  My contribution to the EndoMapper dataset is recording the calibration se-

quences, obtaining the photometric calibration parameters of the endoscopes and providing a full photometric calibration software library. The dataset is publicly available at `https://doi.org/10.7303/syn26707219`.

- J. Rodríguez-Puigvert*, V. M. Batlle*, J. M. M. Montiel, R. Martinez-Cantin, P. Fua, J. D. Tardós, and J. Civera. LightDepth: Single-view depth self-supervision from illumination decline. In *IEEE Int. Conf. on Computer Vision (ICCV)*, pages 21273–21283, 2023.
  The asterisk denotes equal contribution. My main contribution to this work is the photometric loss term and the evaluation on the synthetic dataset, including creating the synthetic dataset and running the experiments.

- V. M. Batlle, J. M. M. Montiel, P. Fua, and J. D. Tardós. LightNeuS: Neural surface reconstruction in endoscopy using illumination decline. In *Int. Conf. on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 502–512. Springer, 2023a.

- R. Iranzo*, V. M. Batlle*, J. D. Tardós, and J. M. Montiel. EndoMetric: Near-light metric scale monocular SLAM. In *Int. Conf. on Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Springer, 2025.
  The asterisk denotes equal contribution. I contributed to the theoretical model and the experiments, especially those involving the EndoMapper dataset.

1.3.2 Patent

- J. Rodríguez-Puigvert, V. M. Batlle, J. D. Tardós, J. M. M. Montiel, J. Civera, R. Martinez-Cantin, and P. Fua. Self-supervised method for obtaining depth, albedo and surface orientation estimates of a space illuminated by a light source, 2023. Patent application PCT/EP2024/066877.
  The invention is currently being utilized under agreement by international industry partners. A PCT application has been filed, with entry into the national phase expected in the coming months.

1.3.3 Open-source software

- V. M. Batlle, J. M. M. Montiel, and J. D. Tardós. EM Dataset: Photometric calibration, 2023b. URL `https://github.com/endomapper/EM_Dataset-PhotometricCalibration`.
  This repository contains the photometric calibration library for the EndoMapper dataset. See more details in Appendix A.

Chapter 2

Photometric Model for 3D Reconstruction in Endoscopy

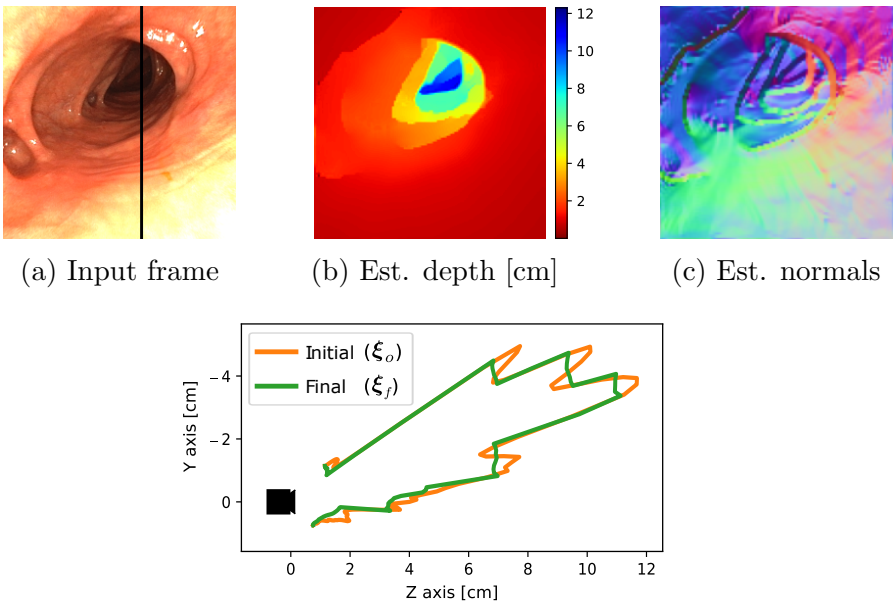
2.1 Introduction

With the goal of improving the efficiency and effectiveness of routine diagnostic and medical intervention procedures, there is a growing research interest in extending augmented reality and autonomous navigation to the human body. These advances will need to accurately solve localization and mapping from visual sensors.

Simultaneous Localization and Mapping (SLAM) with stereo (Mur-Artal and Tardós, 2017) and visual-inertial (Campos et al., 2021) cameras already provide great accuracy for multi-view reconstruction. In most medical endoscopy applications, space limitations restrict to monocular cameras. With monocular vision, the real scale of the environment cannot be observed, so potential applications are limited to up-to-scale reconstructions which usually suffer from scale drift problems, specially in deforming environments (Lamarca et al., 2020).

However, the interior of the human body is an example of an artificially illuminated environment, where the light source is controlled and linked to the camera movement. Our goal is to take advantage of the illumination to reconstruct 3D scenes, obtaining photometric stereo information by considering a light-camera pair.

The main contributions of this chapter are: (1) a simple photometric model for the endoscope light and camera, (2) a photometric calibration method that does not require a Lambertian pattern and can be carried out in-place on a hospital setting, and (3) the first method capable of reconstructing the geometry of the human colon from a single view, only from the illumination of a calibrated conventional monocular endoscope (see Figure 2.1). This would solve the scale-drift problem in monocular SLAM, and can also provide real-scale maps when the albedo of the surface and the endoscope's auto-gain are known.



(a) Input frame (b) Est. depth [cm] (c) Est. normals

(d) Cross-section along black line

Figure 2.1: Depth estimation on an *in-vivo* human colonoscopy

2.2 Related work

Recent results in single-view depth estimation using deep convolutional networks (Godard et al., 2019) open the possibility of designing accurate SLAM systems from monocular cameras, specially hybrid approaches (Yang et al., 2020) which combine deep learning with traditional methods. Previous work demonstrates that, using a depth estimation network, it is possible to perform scale-aware monocular SLAM, obtaining almost the same accuracy as with stereo, and eliminating scale drift (Li et al., 2018a; Tiwari et al., 2020; Campos and Tardós, 2022).

A first attempt to apply these methods to endoscopy sequences achieves real-scale reconstructions with good accuracy (Recasens et al., 2021). However, these methods require stereo supervision to learn how to predict the true size. Thus, today its application in colonoscopy is limited by the impossibility of acquiring stereo images of the human colon.

Other authors focus on the study of photometry to obtain dense and semi-dense reconstructions of outdoor and indoor 3D scenes (Newcombe et al., 2011; Engel et al., 2017). They assume constant illumination, usually ambient light, ignoring any change in lighting.

In contrast, inside the human body, the illumination is controlled and light moves together with the camera. Recent work (Modrzejewski et al., 2020; Hao et al., 2020a) shows that changes in lighting, instead of being ignored, can be used to our benefit, obtaining dense reconstructions from monocular sequences.

2.2.1 Lighting model

Previous works propose a lighting model for their working environment. Specifically, they model light emission, interaction with surfaces, and capture by the camera. So far, the complexity of these *ad-hoc* models required them to be calibrated and tested in laboratory environments. In this chapter, we propose a simplified model that allows easier calibration without the need for Lambertian patterns.

Modrzejewski et al. (2020) do a thorough work on analyzing various light source models. Their Spot Light Source (SLS) model offers a good compromise between complexity and accuracy. We adopt a similar approach, but with the aim of modeling the multiple light sources of the endoscope as a single virtual light. This leads us to a generic model in which our virtual light is located at the camera's optical center.

Hao et al. (2020b) calibrate the light emission separately by means of a plane mirror. Conversely, we propose a joint calibration method, which also allows easy estimation of camera geometry and photometry at the same time.

A common approach (Visentini-Scarzanella and Kawasaki, 2015; Modrzejewski et al., 2020) consists of assuming Lambertian surfaces, both during calibration and during reconstruction inside the human body. However, we show that this causes bias in the calibration when the real surface is not perfectly Lambertian. If not corrected, this error propagates to the 3D reconstruction. In contrast, our calibration considers non-Lambertian properties, giving results that are not affected by the calibration pattern.

2.2.2 Reconstruction

In photometric stereo, the discussed light model can be used to obtain a dense depth map of the scene. The main discrepancy between the different approaches lies in their corresponding reconstruction method.

Modrzejewski et al. (2020) propose an initial multi-view reconstruction followed by a photometric optimization, where a regularization term tends to favor smooth planar surfaces. Crucially, the multi-view method requires a rigid environment. In contrast, our method is based only on lighting, being able to reconstruct the environment from a single view. In addition, the geometry of the human colon, with numerous discontinuities, is far from planar. By considering this in our regularization, we can reconstruct its complex shape.

Hao et al. (2020a) focus on specular highlights, where their method achieves the best accuracy. However, the accuracy of their reconstruction decreases for the rest of the continuous surface. Unlike them, we perform a global optimization, considering each point equally, which does not require the surface to be continuous.

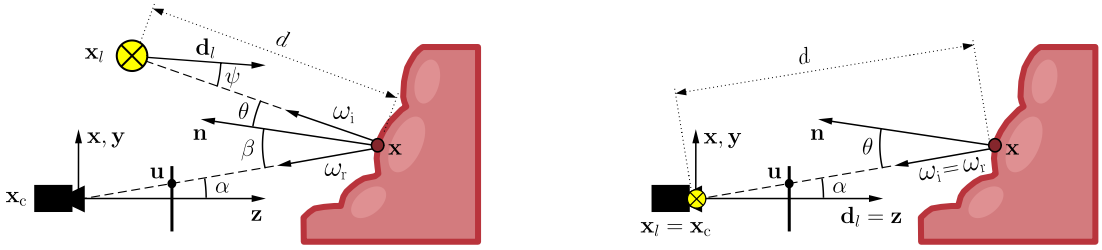


Figure 2.2: **Left:** General photometric model. **Right:** Simplified photometric model. We assume a virtual light is located at the camera optical center and the light's principal direction is at the camera forward vector.

To the best of our knowledge, previous work on dense photometric reconstruction on endoscopy (Okatani and Deguchi, 1997; Collins and Bartoli, 2012b; Hao et al., 2020a; Modrzejewski et al., 2020) has been validated on nearly planar scenes, without discontinuities. Focusing on colonoscopy, Parot et al. (2013) provide experimental validation on phantoms. Instead, we demonstrate that our method can recover for the first time the tubular topology of a human colon, form a single *in-vivo* video frame, preserving the anatomical folds of the intestine, known as haustra.

2.3 Endoscope model

This chapter presents a photometric approach to the problem of monocular 3D reconstruction during medical endoscopy. This approach considers the geometric model of image formation and the photometric model of light transport (Figure 2.2).

2.3.1 Geometric model

An endoscope camera is designed to cover a wide view angle. Thus, we adopt the approach of Kannala and Brandt (2006) to model the fisheye lens, with four projective parameters f_x, f_y, C_x, C_y, and four distortion coefficients k_{1-4}. We denote the optical center of the camera as $\mathbf{x}_\mathrm{c}$.

2.3.2 General photometric model

The illumination system mounted on an endoscope usually consists of one or more small lights. Given its small size, these lights are usually modeled as punctual lights (Modrzejewski et al., 2020; Hao et al., 2020a). Thus, radiance L_i coming from light center $\mathbf{x}_l$ to a point $\mathbf{x}$ in a surface, is subject to the inverse-square law:

$$L_\mathrm{i}(\mathbf{x}) = \mu(\mathbf{x}) \frac{\sigma_0}{\|\mathbf{x} - \mathbf{x}_l\|^2} \tag{2.1}$$

In most endoscopes light is transmitted to the tip using optical fiber. Therefore, the amount of light emitted in each direction of space is not uniform. We model this behavior with a light spread function $\mu(\mathbf{x})$, that can be specified by fixing a principal direction $\mathbf{d}_l$, over which maximum radiance σ_0 is emitted, and assuming a radial cosine fall-off (Hao et al., 2020a). We decide to modulate this decay by adding the cosine exponent k as a parameter:

$$\mu(\mathbf{x}) = \cos^k \psi, \quad \psi = \langle \mathbf{x} - \mathbf{x}_l,\ \mathbf{d}_l \rangle \tag{2.2}$$

When light reaches a surface, most of it will be reflected, going out in different directions depending on the material properties. The Bidirectional Reflectance Distribution Function (BRDF) $f_\mathrm{r}(\omega_\mathrm{i}, \omega_\mathrm{r})$ defines how light is reflected at an opaque surface. Usually, this behavior depends on the incoming ω_i and outgoing ω_r direction of the light ray with respect to the normal $\mathbf{n}$ of the surface at that point. The inclination of the incident ray modifies the area of the projection of the solid angle on the surface, depending on the cosine of its angle with the normal. Thus, the reflected radiance is:

$$L_\mathrm{r}(\mathbf{x}, \omega_\mathrm{r}) = L_\mathrm{i}(\mathbf{x}, \omega_\mathrm{i})\ f_\mathrm{r}(\omega_\mathrm{i}, \omega_\mathrm{r}) \cos\theta \tag{2.3}$$

where $\theta = \langle \omega_\mathrm{i}, \mathbf{n} \rangle$ and, for our case, ω_i is the direction to the light source and ω_r points to the camera (see Figure 2.2).

Light reaching the camera is affected by a set of factors. The capture system usually introduces attenuation on the received radiance. Natural vignetting tends to approximate to $\cos^4 \alpha$, where α is the off-axis angle between the ray direction and the camera forward $\mathbf{z}$ (Szeliski, 2010). Mechanical vignetting is not easy to model theoretically. Therefore, vignetting is usually empirically approximated (Engel et al., 2016). We assume radial attenuation from the camera's forward vector, by modeling the decay with a k' exponent on a cosine function:

$$V(\mathbf{x}) = \cos^{k'} \alpha, \quad \alpha = \langle \mathbf{x} - \mathbf{x}_\mathrm{c},\ \mathbf{z} \rangle \tag{2.4}$$

Endoscope cameras might automatically adjust some parameters, such as exposure time or signal amplification. In video streams, this is controlled by an automatic gain control (AGC) logic. We assume this auto-gain usually acts as a multiplying factor g_t at each t-th time instant. In order to increase the perceived dynamic range, cameras map the captured values through a gamma function with a common value of $\gamma = 2.2$ that does not change over time.

Our complete photometric model considers all concepts introduced above, as a combination of light, surface and camera effects:

$$\mathcal{I}(\mathbf{x}) = \left(\frac{\mu(\mathbf{x})\ \sigma_0}{\|\mathbf{x} - \mathbf{x}_l\|^2}\ f_\mathrm{r}(\omega_\mathrm{i}, \omega_\mathrm{r})\ \cos\theta\ V(\mathbf{x})\ g_\mathrm{t} \right)^{1/\gamma} \tag{2.5}$$

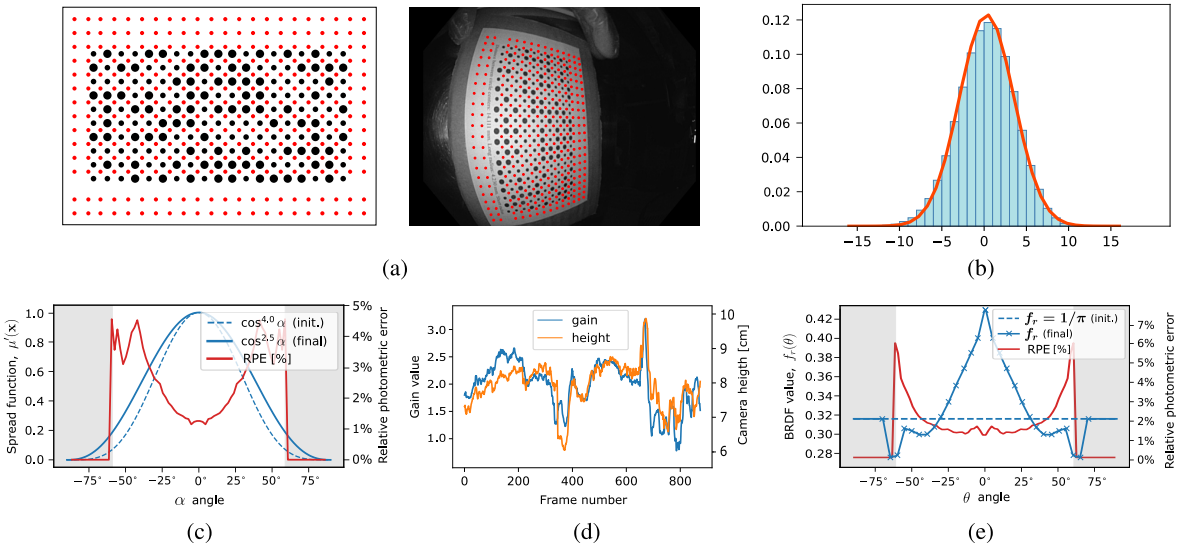


Figure 2.3: Sampling the Vicalib pattern: (a) Red marks correspond to each $\boldsymbol{x}_j$ sampled point. Photometric calibration results: (b) Photometric errors of the calibrated model are close to a Gaussian distribution with a mean of 0.3 and std. of 3.2 gray levels. (c) Joint attenuation caused by light spread function $\mu(\mathbf{x})$ and camera vignetting $V(\mathbf{x})$. (d) Estimated auto-gain factors over the calibration video. (e) Non-Lambertian BRDF for the paper sheet used for calibration.

2.3.3 Simplified photometric model

The presented model takes into account only one light source. Each additional light source must be modeled independently, in a similar way, but adding complexity to the model. Moreover, the characteristics of each endoscope version vary slightly. Commonly we find two or three optical fiber guides, which conduct the light to different points on the tip of the endoscope.

Instead of the costly process of modeling the details of each specific hardware, we propose a simplification based on encapsulating the joint effects of all the light points into a single virtual light source (see Figure 2.2). We observe that these light points are usually distributed fairly symmetrically around the endoscope camera. Therefore, we decided to place the virtual light source at the optical center of the camera, i.e. $\mathbf{x}_l = \mathbf{x}_\mathrm{c}$ and we align light's principal direction with the camera forward, i.e. $\mathbf{d}_l = \mathbf{z}$.

In this new set-up, camera's vignetting and virtual light's spread function are coupled, i.e. $\psi = \alpha$. Thus, we model the effect of both functions jointly, as:

$$\mu'(\mathbf{x}) = \mu(\mathbf{x})\, V(\mathbf{x}) = \cos^k \alpha \tag{2.6}$$

Regarding surface reflectance, now incoming ω_i and reflected ω_r directions match. Thus, the domain of the BRDF can be simplified. We will consider the incident angle θ of the light on the surface, such that the BRDF is simply $f_\mathrm{r}(\theta)$.

In most endoscopes auto-gain logic is unknown. Therefore, g_t values are coupled with the absolute radiance σ_0, so that their effects cannot be separated. Consequently, we fix the σ_0 parameter to an arbitrary value and estimate the relative auto-gain changes.

Finally, we obtain a simplified photometric model, which is parameterized according to the unknowns we want to estimate for our endoscope:

$$\mathcal{I}\left(\mathbf{x},\, k,\, g_\mathrm{t},\, \gamma\right) = \left(\frac{\mu'(\mathbf{x},\, k)}{\|\mathbf{x} - \mathbf{x}_l\|^2}\; f_\mathrm{r}(\theta)\;\cos\theta\; g_\mathrm{t} \right)^{1/\gamma} \tag{2.7}$$

2.4 Endoscope calibration

Geometric and photometric calibration is performed with a small Vicalib (Heckman et al., 2016) pattern printed on a white paper sheet of 5.61×9.82 cm. From a video captured with the endoscope, geometric parameters are obtained by processing 1 out of 20 frames using the Vicalib software.

Focusing on photometry, we propose an optimization problem that aims to minimize the photometric error between the empirical data I and our model. For that, we select a set of j sample points, uniformly distributed along the white areas of the pattern (see Figure 2.3a). Then, the photometric loss function is computed on every visible $\boldsymbol{x}_\mathrm{j}$ point on each t-th frame of the video, using Huber function ρ to be robust against spurious:

$$\{k, g_\mathrm{t}, \gamma \mid \forall \mathrm{t}\}^* = \underset{k,\, g_\mathrm{t},\, \gamma}{\operatorname{argmin}} \sum_{\mathrm{j,t}} \rho\left(I_\mathrm{jt} - \mathcal{I}\left(\boldsymbol{x}_\mathrm{j}, k, g_\mathrm{t}, \gamma\right)\right) \tag{2.8}$$

2.4.1 Results

From a video resolution of 1440×1080 px at 30 frames per second, the calibration obtains values for the geometric parameters $f_x = 717.21$ px, $f_y = 717.48$ px, $C_x = 735.37$ px, $C_y = 552.80$ px and the four distortion coefficients k_{1-4} are $[-0.13893,\ -1.2396E-03,\ 9.1258E-04,\ -4.0716E-05]$. These lead to a reprojection error in RMSE of 0.5288 px.

Regarding the photometric calibration results, the optimization converges to $k = 2.5$, $\gamma = 2.2$ and estimates auto-gain g_t values ranging from 1 to 3. We observe that the final spread is wider than the natural vignetting $\cos^4\alpha$ (see Figure 2.3c). This is consistent with the illumination system of our endoscope, where three real light sources result in a widening of our virtual light's spread. Validation results show a small relative error of 1% in the center, i.e. 0°. However, error grows towards the

edges, reaching 4.5% at 40° and when $\alpha > 60°$ the function remains unsampled in our calibration data (shaded area in Figure 2.3c and e).

Automatic gain cannot be evaluated with test data, because each image has a different gain factor. Instead, we can see that the estimated gain factor follows a continuous progression over time along the calibration sequence (see Figure 2.3d). Moreover, the gain value for each frame seems to be closely related to the distance from the camera to the illuminated surface. That is, when the camera is closer to the pattern, light is more intense, and the endoscope applies a lower gain value.

In addition, we observed that modeling the paper reflectance with a Lambertian BRDF $f_{\mathrm{r}}(\theta) = 1/\pi$ led to a biased calibration. Therefore, we decided to also optimize the value of the BRDF for fifteen values of the θ angle and apply linear interpolation in the rest of the domain of the function (see Figure 2.3d). The estimated BRDF for the paper sheet shows specular behavior when the camera is close to the perpendicular to the surface. This results in a peak in the reflectance when the θ angle is near zero.

in The result of the calibration allows us to estimate the gray level of a pixel with a standard deviation of 3.2 levels. The distribution of errors is unbiased (see Figure 2.3b). Moreover, the new estimated BRDF is an isolated component of the model. Therefore, when we want to apply our calibration in the interior of the human body, we can replace this BRDF with that of the human colon, and the rest of the calibrated parameters remain valid.

2.5 Depth estimation

Given the calibrated endoscope photometric model and a single endoscope image, our goal is to estimate depth and surface normal for each imaged 3D point. We consider the following assumptions:

- Similarly to Modrzejewski et al. (2020), we assume that human tissue can be approximated by a Lambertian material if specular highlights are masked or treated as spurious. For this, we propose an automatic method for highlight detection and inpainting.

- In addition, given the weak texture of the colon tissue, the surface albedo k_d is measurable and considerably constant. In our experiments we set $f_{\mathrm{r}}(\theta) = k_d/\pi$.

- The imaged surfaces are smooth, except at occasional discontinuities. This allows us to approximate differential changes of the surface by a tangent plane.

Based on DTAM method (Newcombe et al., 2011), we approach the estimation of

a depth map as an optimization problem, that minimizes an energy function:

$$E_{\boldsymbol{\xi}} = \int_{\Omega} \Big\{ C(\mathbf{u}, \boldsymbol{\xi}(\mathbf{u})) + \lambda R(\mathbf{u}, \boldsymbol{\xi})) \Big\} d\mathbf{u} \tag{2.9}$$

where

- $\mathbf{u} \in \Omega$ are coordinates on the image,
- $\boldsymbol{\xi} : \Omega \to \mathbb{R}$ is the depth map,
- $C(\,)$ is a photometric cost function,
- $R(\,)$ is a regularization cost,
- $\lambda \in \mathbb{R}^+$ adjusts the regularization weight.

2.5.1 Photometric cost function

DTAM assumes ambient light on the scene and uses a cost function based solely on camera geometry and brightness constancy. However, the illumination during endoscopy varies with camera movement. Consequently, we replace the original cost function with a novel cost function based on our photometric endoscope model:

$$C(\mathbf{u}, d) = \rho \left(I(\mathbf{u}) - \mathcal{I} \left(\pi^{-1} \left(\mathbf{u}, d \right) \right) \right) \tag{2.10}$$

where

- d is the Euclidean distance to the world point,
- $\pi^{-1}(\,)$ is the camera unprojection model,
- $\mathcal{I}(\,)$ is our calibrated endoscope photometric model (2.7),
- $I : \Omega \to \mathbb{R}^+$ denotes the actual pixel intensity,
- $\rho(\,)$ is the Huber robust cost function.

2.5.2 Normal estimation from a depth map

The photometric model of a scene $\mathcal{I}$ is influenced by both the distance to the points (inverse-square law) and the surface normal (cosine term). However, surface normal is directly related to depth variations. Therefore, both parameters should not be optimized separately. Instead, given the local planarity assumption, we can calculate the normal of a point from the estimated depth map (Hinterstoisser et al., 2011).

Thanks to this relationship, we keep the depth map as the only unknown variable of the problem. However, it should be noted that this method is influenced by spurious data, especially at surface discontinuities. Therefore, in these areas, we expect some localized errors.

2.5.3 Smoothness regularization

The defined cost function is trying to find three unknowns per pixel (d, n_θ, n_φ) from one intensity measurement (I). In order to solve the problem's ill-posedness, DTAM proposes a regularization term that penalizes local depth variations, except at points where the luminosity gradient is large, which usually correspond to surface discontinuities:

$$R(\mathbf{u}, \boldsymbol{\xi}) = g(\mathbf{u}) \|\nabla \boldsymbol{\xi}(\mathbf{u})\|_\epsilon \tag{2.11}$$

Thus, $\|x\|_\epsilon$ Huber norm with $\epsilon \approx 1.0e^{-4}$ works as total variation (TV) regularizer and $g(\mathbf{u})$ reduces the regularization strength at high gradient points. Thanks to these two terms, (d, n_θ, n_φ) are now constrained by the pixel's neighborhood, and at the same time the discontinuities of the colon can be preserved.

However, this might not be the best regularizer in the colon's tubular geometry. The first derivative $\nabla \boldsymbol{\xi}$ always favors zero changes along the depth map. So the reconstruction will tend to a plane parallel to the camera. Instead, similarly to Modrzejewski et al. (2020), we use the second-order derivative $\nabla^2 \boldsymbol{\xi}$ to impose smoothness, although we continue to allow discontinuities.

2.5.4 Depth map representation

DTAM formulates its depth map $\boldsymbol{\xi}_{1/z}$ as the inverse distance in the Z-axis. This decision is appropriate for a multi-view-based problem, as the pinhole projection model depends directly on this variable. Instead, we are faced with a single-view problem. In our case, the photometry is quadratically dependent on the inverse of the Euclidean distance, i.e. $\mathcal{I} \propto 1/d^2$.

Therefore, we will compare the previous formulation with two new variants of the depth map, such that

$$\boldsymbol{\xi}_{1/z} = \frac{1}{z}, \qquad \boldsymbol{\xi}_d = d, \qquad \boldsymbol{\xi}_{1/d} = \frac{1}{d} \tag{2.12}$$

2.5.5 Initial solution

We make the optimization method start from an initial solution, where we assume all surface normal vectors pointing towards the camera optical center.

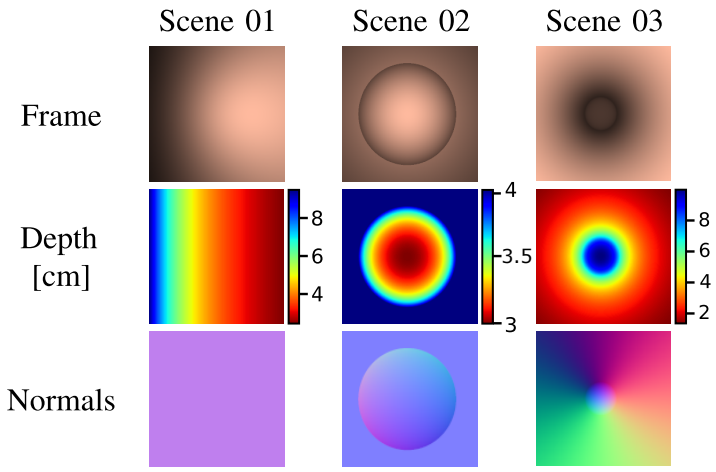


Figure 2.4: Data generated for our simple geometry dataset. **Top:** Frame simulated with our model as in (2.7). **Middle:** Ground-truth Z-depth map. **Bottom:** Ground-truth normal map, represented in color space $(R, G, B) = ([n_x, n_y, n_z] + 1)/2$.

From the calibrated photometric model, we revert the effects of light spread function, Lambertian BRDF, as well as known camera gain and gamma correction:

$$I_{\mathrm{c}}(\mathbf{u}) = \frac{I(\mathbf{u})^{\gamma}}{\mu'(\mathbf{x}) \cdot f_{\mathrm{r}}(\theta) \cdot g_{\mathrm{t}}} = \frac{\cos\theta}{d^2} \tag{2.13}$$

where I_{c} is a *canonical intensity value*, which is obtained after compensating all mentioned parameters that influence image formation. Note that, when a surface normal points towards the camera, the θ angle is zero. Therefore, by solving for d in the above equation, we get an initial solution

$$d_o(\mathbf{u}) = I_{\mathrm{c}}(\mathbf{u})^{-1/2} \tag{2.14}$$

The closer the actual θ is to zero, the closer this initial solution is to the real depth.

2.6 Experimental results

In this section, we first conduct some simple experiments to determine the best smoothness regularizer and depth map variant. Then, we check the accuracy of our depth estimation method with a photo-realistic simulation of a human colon. Finally, we test our method in a real in-vivo colonoscopy image. For nonlinear optimization we use a trust region algorithm with numerical Jacobian matrix computation.

2.6.1 Simple geometry dataset

The first experiment consists of a simple simulation, based on our photometric model as in the equation (2.7). As shown in Figure 2.4, we simulate a rotated plane, a curved surface, and a tubular geometry.

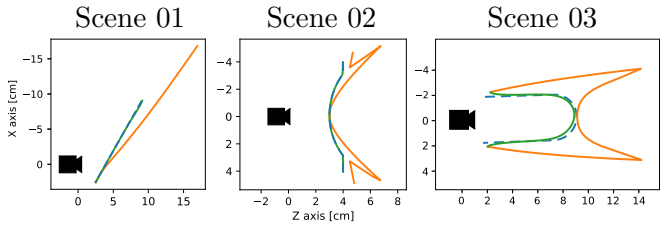


Figure 2.5: Cross-section along Y-axis (middle of the image) of actual surface (GT, dashed blue), initial solution ($\boldsymbol{\xi}_o$, orange) and final optimized depth map ($\boldsymbol{\xi}_f$, green). Our estimation matches the ground-truth.

Table 2.1: Accuracy on our simple geometry dataset

Scene	$\boldsymbol{\xi}$	**Reg.**	**# iter.**	**Depth error [mm]**		**Depth error [%]**		**Normals error [deg]**	
				Mean	**Median**	**Mean**	**Median**	**Mean**	**Median**
01	$1/z$	∇	1 148	1.0	**<0.1**	0.90	**0.04**	1.19	0.20
		∇^2	**73**	**0.3**	**<0.1**	**0.32**	0.09	**0.62**	**0.18**
02	$1/z$	∇	102	0.2	0.2	0.35	0.37	1.00	0.50
		∇^2	**64**	**0.1**	**0.1**	**0.25**	**0.21**	**0.95**	**0.39**
03	$1/z$	∇	>5 550	3.0	2.1	7.30	6.73	12.17	9.44
		∇^2	>1 500	**1.9**	**1.8**	5.83	5.31	**11.07**	**8.00**
	d	∇^2	301	**1.9**	**1.8**	5.85	**4.99**	12.92	9.15
	$1/d$	∇^2	**78**	**1.9**	**1.8**	**5.78**	5.21	11.55	8.30

Table 2.1 presents the results of this experiment. We conclude that the regularizer of the second derivative is better for accuracy (see in Figure 2.5 the best result for each scene) and is also faster in convergence.

Regarding the depth map variants, in a tubular geometry, inverse Euclidean distance $\boldsymbol{\xi}_{1/d}$ performs better than the alternatives. The corresponding results in Table 2.1 show much faster convergence with similar accuracy.

2.6.2 Simulated colon dataset

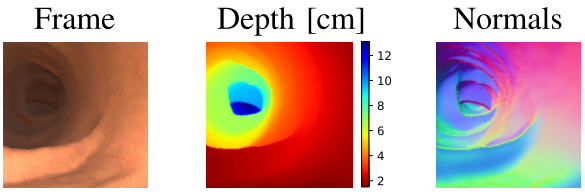
In this experiment, we validate our method on a frame of a photo-realistic dataset (Rau et al., 2023). This dataset simulates an endoscopy procedure based on a real CT scan of a human colon. The simulation includes effects more similar to those found in a real environment, such as richer textures and ambient light caused by secondary reflections, that are not considered in our model.

This input frame comes from an endoscope without distortion or vignetting and in which the light spread is homogeneous (see Figure 2.6a). We also know the average albedo of the surface and the gain of the endoscope. The results of this experiment are shown in the Table 2.2.

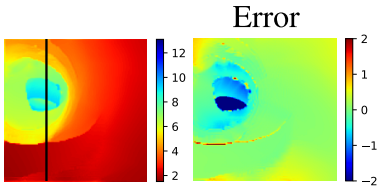
As in the previous case, we see that $\boldsymbol{\xi}_{1/d}$ is the best variant of the method, providing a good estimation (see Figure 2.6b) with less than 3 mm error on average. However, on

Table 2.2: Accuracy on a simulated colon (Rau et al., 2023)

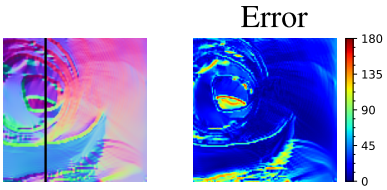
$\boldsymbol{\xi}$	Reg.	# iter.	Depth error [mm]		Depth error [%]		Normals error [deg]	
			Mean	Median	Mean	Median	Mean	Median
$1/z$	∇	44 500	5.3	2.3	10.60	7.41	30.63	23.77
	∇^2	8 900	5.1	4.0	15.09	11.86	36.06	29.26
d	∇	20 000	3.3	**1.6**	7.90	**4.98**	**26.21**	**18.75**
	∇^2	44 500	3.8	2.0	9.57	6.37	32.00	23.56
$1/d$	∇	44 500	**2.8**	**1.6**	**7.32**	5.01	27.89	19.69
	∇^2	**5 400**	5.1	4.0	15.16	12.05	35.86	28.89



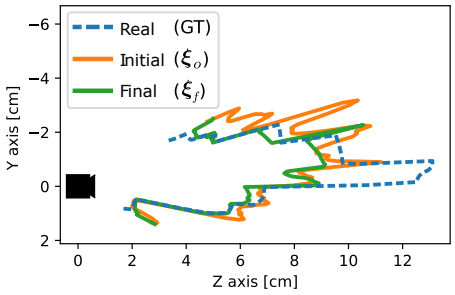
(a) Ground-truth



(b) Depth estimation [cm]



(c) Normals estimation [deg]



(d) Cross-section along black line

Figure 2.6: Results on the simulated colon dataset (Rau et al., 2023). Using variants $\boldsymbol{\xi}_{1/d}$ and $\nabla\boldsymbol{\xi}$. They show good accuracy for the estimated Z-depth map and the corresponding normals.

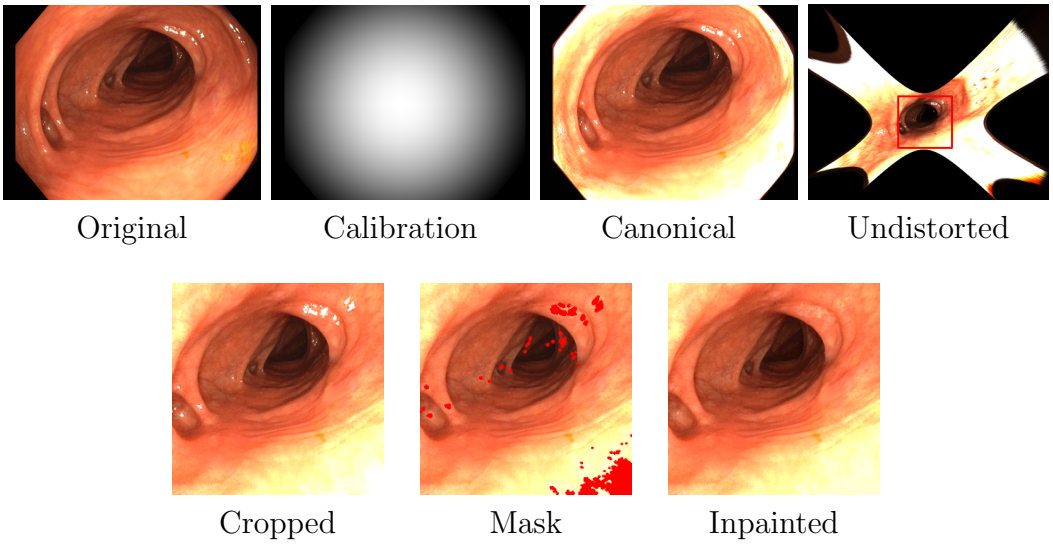


Figure 2.7: Results on real colon dataset (Azagra et al., 2023) sequence Seq_003. We apply our photometric and geometric calibration and we perform highlight inpainting.

this photo-realistic simulation, the ∇^2 regularizer obtains lower accuracy. The second derivative is less robust to noise, such as that introduced by surface texture, which causes the albedo to be not perfectly uniform.

In addition, the photo-realistic simulator introduced a fog effect in areas far away from the camera. This increases the intensity of distant pixels. As a result, we cannot reconstruct the deepest part of the colon, from 10 to 12 cm (see Figure 2.6d). Therefore, the median error of 1.6 mm is considerably lower than the mean, which is influenced by those spurious.

2.6.3 Real colon dataset

Our method is validated on a real image from the EndoMapper colonoscopy dataset (Azagra et al., 2023). This image corresponds to a real human colon and was acquired *in-vivo* during a medical procedure, with the endoscope we calibrated in section 2.4.

We take a single image (see Figure 2.7). First, we compensate for calibrated vignetting and light spread to obtain a frame with canonical illumination, which we undistort and crop. Then, we perform automatic highlight detection and inpainting. This leads to a frame similar to the one of the previous simulated dataset.

Figure 2.1 shows the reconstruction provided by our method. The estimated scale is arbitrary, as we do not have data about the automatic gain. Moreover, the EndoMapper dataset does not provide ground-truth information for comparison. Nevertheless, qualitatively, the result we have obtained properly reconstructs the tubular topology of the colon and also recovers notably the shape of the haustra.

2.7 Conclusions

This chapter proposes a photometric stereo method that is able to reconstruct for the first time the geometry of the human colon using only the illumination on real monocular endoscopy procedures. We can recover the true scale of the environment if the surface albedo and the endoscope's auto-gain are known. The latter is set by the manufacturer of the hardware and the albedo could be reasonably estimated in the future since it is mostly uniform along the colon.

Our method obtains reconstructions with a mean accuracy below 3 mm on simulated data and is able to reconstruct the tubular geometry on a real colon, where it preserves the discontinuities at the colon's haustra.

In addition, a calibration process is designed to suit a medical endoscope. Our experiments show that we are able to model a real endoscope with an error of 3 gray levels. This allows us to conclude that our model, based on a virtual light source, offers a good compromise between accuracy and ease of calibration in a real environment.

Currently, depth estimation works in an off-line mode, but it allows us to overcome the lack of 3D perception inherent in monocular camera systems. In this way, for example, our method could constitute a new source of self-supervision for learning depth estimation without the need for stereo.

In conclusion, these results open the door to future work in real-time SLAM and autonomous navigation in the colon, solving scale drift and allowing true scale maps.

Chapter 3

Single-View Depth Self-Supervision from Illumination Decline

3.1 Introduction

Minimally invasive medical procedures such as gastroscopies, colonoscopies and bronchoscopies rely on endoscopes that should be as small as possible. As a result, they usually house a single camera and several light points, but neither depth nor stereo cameras. 3D reconstruction is relevant in endoscopies, as it may unlock several functionalities such as the accurate estimation of the size and shape of tumors. However, both single- and multi-view depth estimation methods present significant challenges in this domain. The lack of sufficient depth annotated data hinders the use of supervised depth learning. The presence of fluids that either obscure the view or generate specularities, the sudden illumination changes, the paucity of texture and the surface deformations hamper multi-view methods both for self-supervising deep networks and for geometry estimation. Real in-body textures and fluids are hard to simulate realistically, and the synthetic-to-real gap may be large.

In this work, we propose a novel approach to depth in endoscopies that overcomes all the above challenges related to depth supervision, multi-view estimation and synthetic-to-real gaps. Our key insight is that, by exploiting a key property of endoscopic imagery, we can provide strong depth self-supervision signals from just one view. In endoscopes, the light source is rigidly located next to the camera and is close to the surface to be reconstructed. As a result, unlike in traditional shape-from-shading (SfS), points with the same albedo are imaged darker the further they are, being the decrease of intensity a function of the square distance to the light source. To exploit this, we introduce a deep network, as depicted by Figure 3.1, that estimates depths and albedos from the image, infers normals from depths, and then renders an image while taking into account the attenuation factor due to the distance between the light source and the

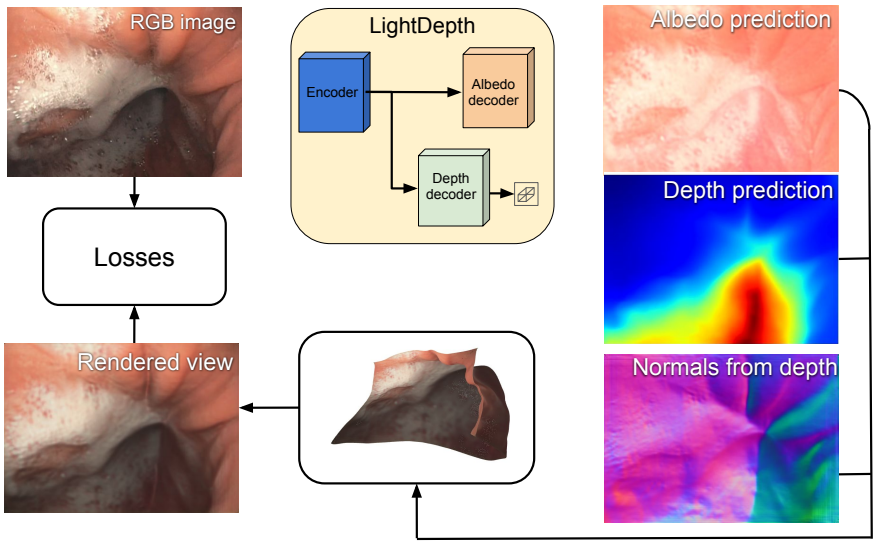


Figure 3.1: **Single-view depth self-supervision in LightDepth.** A two-headed deep network predicts albedo and depth from a single image and estimates surface normals from predicted depths. These are used to render a new image, that takes into account illumination decline and the endoscope's photometric calibration, and can be compared to the original one. Minimizing the difference between the original and rendered images is used at training time to compute the network weights and at inference time to refine the depth predictions.

surface. At training time, we minimize the difference between the original and rendered images. This enforces consistency of the depths, normals, and albedos and provides the required self-supervision without depth annotations. At inference, we use our trained network to predict depth from a RGB image and then, as our method is totally self-supervised, we can perform test-time refinement (TTR) for every monocular image, minimizing the difference between the input and rendered views, further refining the predicted depths. Our quantitative evaluation on a phantom colon dataset, where ground-truth is available, shows that our *self-supervised* approach delivers results that are very close to that of the best supervised one, and significantly superior to that of multi-view self-supervision and synthetic-to-real transfer methods. Crucially, we show quantitatively that our method keeps working on real data, for which there is no ground-truth data that can be used for training and self-supervised alternatives underperform. The main specific contributions that led to such results are 1) the inclusion of illumination decline and the endoscope's photometric calibration in the rendering equation, which provides a strong supervisory signal, and 2) a single-view self-supervised method using such renders, including two-headed network architectures LightDepth U-Net and LightDepth DPT (see details in Figure 3.3) that can be trained in large colonoscopy datasets without requiring ground truth labels and even further refined online in the test views.

3.2 Related work

Generic Single-view Depth Estimation. It has enjoyed a renaissance after the seminal work by Eigen et al. (2014), which demonstrated the effectiveness of deep neural networks for supervised pixel-wise depth regression in natural images. Subsequent research efforts have made contributions in many different directions. To name a few, network architectures evolved to fully convolutional in Laina et al. (2016) and more recently to transformers (Ranftl et al., 2021; Bhat et al., 2021; Li et al., 2024). Some of those works (Bhat et al., 2021; Li et al., 2024) also discretize the continuous depth space into bins and formulate the problem as an ordinal regression, as in Fu et al. (2018). Other advances include interpretability (van Dijk and de Croon, 2019), uncertainty quantification (Poggi et al., 2020; Rodríguez-Puigvert et al., 2022), and modeling camera intrinsics (Fácil et al., 2019; Gordon et al., 2019).

All those approaches are supervised and require depth ground-truth data, which can be difficult and expensive to acquire. Self-supervised methods seek to overcome this limitation and reduce the need for ground-truth data, often by exploiting multi-view photometric consistency (Godard et al., 2017; Zhou et al., 2017, 2018a; Yang et al., 2018; Godard et al., 2019; Johnston and Carneiro, 2020). This also enables depth refinement at test time (Chen et al., 2019c; Tiwari et al., 2020; Luo et al., 2020; Shu et al., 2020; Watson et al., 2021; Izquierdo and Civera, 2023). Unfortunately, this kind of supervision can be noisy, due to inaccuracies in the camera motion estimation, perspective distortions, occlusions or non-Lambertian effects, among others. As result, state-of-the-art self-supervised methods typically suffer from significantly larger inaccuracies than supervised ones. By contrast, our approach avoids these sources of errors and delivers accuracies that are close to those of supervised techniques.

Endoscopic Single-view Depth Estimation. Single-view depth estimation has been extensively studied for endoscopic purposes. Visentini-Scarzanella et al. (2017) used CT renderings for depth supervision in bronchoscopies. However, CT scans in particular and ground-truth depth data in general are very rare in endoscopy, which makes self-supervision a quasi necessity. Many works explore multi-view integration (Luo et al., 2019; Xu et al., 2019; Huang et al., 2021) combined with tracking and SLAM pipelines (Recasens et al., 2021; Ozyoruk et al., 2021; Ma et al., 2021). Others propose video-based training schemes (Freedman et al., 2020; Karaoglu et al., 2021; Hwang et al., 2021). Unfortunately multi-view self-supervision is even more challenging in endoscopy than in other areas due to the presence of deformations and weak texture.

Due to the specificity of the domain, synthetic to real transfer has also been extensively explored. For example, in Shen et al. (2019) a conditional GAN is used for

depth recovery while integrating SLAM and multi-view inputs. In Chen et al. (2019b), a depth network is trained with synthetic images of a simple colon model and fine-tuned with domain-randomized photorealistic images rendered from CT scans. Many other works address the domain shift between simulated and real colons (Mahmood et al., 2018; Mahmood and Durr, 2018; Rau et al., 2019; Karaoglu et al., 2021; Cheng et al., 2021; Rodriguez-Puigvert et al., 2022). Learning in supervised and transferring the knowledge using uncertainty (Liu et al., 2019) uses monocular videos and multi-view stereo to provide weak depth supervision. We will show in the results section that our approach yields more accurate results, especially given that our approach to self-supervision allows further refinement of the estimates at inference time.

Shape from Shading (SfS). Depth estimation from a single image can be traced back to the early SfS methods summarized in Zhang et al. (1999) and in particular to traditional shape-from-shading (Horn and Brooks, 1989). However, these older techniques rely on strong assumptions that do not hold in endoscopic imagery: the camera and directional point light model are located at infinity; the reflectance is Lambertian; the albedo is constant, and the surfaces are smooth.

Importantly, lights at infinity result in ill-posed problems (Prados and Faugeras, 2005). By contrast, when the light source is co-located with the camera that is *not* distant from the target surfaces, there is a $1/d^2$ attenuation of pixel intensity with distance d to the surface, which makes the problem well-posed when the albedo is assumed to be constant. Experimental validation that this still holds when the light source is translated with respect to the optical centre is provided in Collins and Bartoli (2012b); Visentini-Scarzanella et al. (2012), but still assuming constant and known albedo. Photometric stereo infers depth capturing several images from the same monocular camera under lights at different locations, but requires endoscopic hardware modifications (Collins and Bartoli, 2012a; Parot et al., 2013; Hao et al., 2020a).

More recently, the topic was revisited by SIRFS (Shape, Illumination, and Reflectance from Shading) (Barron and Malik, 2015) that model the interdependences between shape, illumination and reflectance, and introduces statistical priors on these quantities to disentangle their effects. In subsequent works (Lettry et al., 2018; Li et al., 2020; Sang and Chandraker, 2020; Lichy et al., 2021; Zhang et al., 2022), priors are learned by deep neural networks using supervision, synthetic-to-real or multi-view self-supervision. In contrast, our approach does not require such priors, which makes its deployment easier.

The SfS methods applied to endoscopy require an accurate geometrical and photometrical model of the camera and light source. This can be obtained with calibration (Modrzejewski et al., 2020; Hao et al., 2020b; Batlle et al., 2022; Azagra et al., 2023).

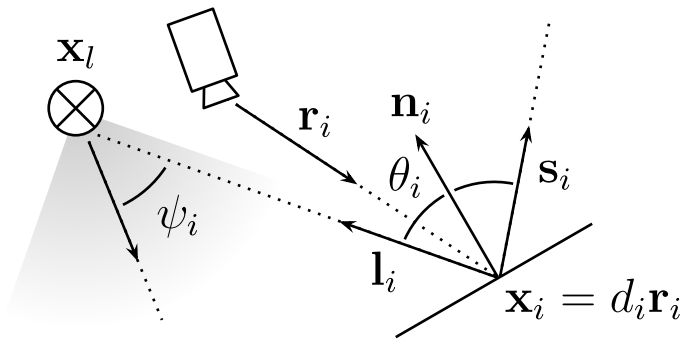


Figure 3.2: Spotlight illumination model, a spotlight source at position $\mathbf{x}_l$ illuminates the surface point $\mathbf{x}_i$. The emission has $R(\psi_i)$ radial fall-off, suffers from an inverse-square decline with $\mathbf{x}_l \rightarrow \mathbf{x}_i$ and attenuates with the incidence angle (θ_i). $\mathbf{l}_i$, $\mathbf{n}_i$, $\mathbf{r}_i$ and $\mathbf{s}_i$ are unit vectors.

3.3 LightDepth

We use a self-supervised single-view approach to train a neural network to predict the albedo, depth, and normals at every pixel of an image so that the image can be resynthesized from these values. As shown in Fig. 3.1, we exploit this property using a dual-branch network that outputs pixel-wise depths and albedos. The normals are estimated analytically from the depths, and, together with the albedos, are used to render images that should be as close as possible to the original ones. At the heart of this approach is the fact that the renderer takes into account light decline as a function to distance to the surface. This is what provides the necessary self-supervisory signal.

3.3.1 Photometric Model

As in Modrzejewski et al. (2020) and Batlle et al. (2022), we model scene illumination as coming from a single spotlight source located at $\mathbf{x}_l \in \mathbb{R}^3$ in the camera reference frame, as depicted by Fig. 3.2. Spotlights usually emit with different intensities in each direction. Hence, we adopt the spotlight model (SLS) of Modrzejewski et al. (2020). For surface point $\mathbf{x}_i$ with off-axis angle ψ_i, we write its radiance as

$$\sigma_{\text{SLS}}(\mathbf{x}_i, \psi_i) = \frac{\sigma_0}{\|\mathbf{x}_i - \mathbf{x}_l\|^2} R(\psi_i) \ , \tag{3.1}$$

$$R(\psi_i) = e^{-\mu(1-\cos(\psi_i))} \tag{3.2}$$

where σ_0 is the maximum radiance and $R(\psi_i)$ is the radial attenuation controlled by a spread factor μ. Note that the light reaching the surface is subject to the inverse-square law and decays with the propagation distance from $\mathbf{x}_l$ to $\mathbf{x}_i$.

Light Decline. In endoscopes, the camera and the light source move jointly in a

dark environment. Hence, the attenuation of the illumination is an indirect indicator of scene depth as seen from the camera. More specifically, for each pixel, we can write the rendering equation

$$\mathcal{I}(d_i, \rho_i, g) = \left(\frac{\sigma_0}{\|d_i \mathbf{r}_i - \mathbf{x}_l\|^2} R(\psi_i) \cos(\theta_i) \ \rho_i \ g \right)^{1/\gamma} , \tag{3.3}$$

where d_i is the depth of the i-th pixel with image coordinates $\mathbf{u}_i$, $\mathbf{r}_i = \pi^{-1}(\mathbf{u}_i)$ is the camera ray such that $\mathbf{x}_i = d_i \mathbf{r}_i$ and $\pi^{-1}(\cdot)$ is the inverse projection model of the camera. θ_i stands for the light's incidence angle with respect to the surface normal $\mathbf{n}_i$, such that, $\cos\theta_i = \mathbf{l}_i \cdot \mathbf{n}_i$. ρ_i represents the albedo of the surface at that point. g denotes the gain applied by the camera and γ is the gamma correction commonly applied by cameras to adapt images to human perception. The resulting $\mathcal{I}(d_i, \rho_i, g)$ is the color captured by the camera.

Our model assumes Lambertian reflections, meaning that the light hitting the surface is scattered equally in all directions. The percentage of reflected light is known as albedo. Specular reflections, which are prevalent in endoscopic images, are not captured by this model but we will consider them in a specific loss that we describe in Section 3.3.2.

Calibration. Each endoscope has different geometric and photometric parameters, the former affecting the inverse project model π^{-1} and the latter impacting both the light position $\mathbf{x}_l$ and spread R. We can estimate these parameters for a particular endoscope by minimizing the reprojection and photometric errors on images of a calibration target, similar to Batlle et al. (2022) and Azagra et al. (2023). In our case, the auto-gain values of the endoscope are not known, so radiance measurements of the camera are unitless. Thus, we arbitrarily set $g = 1$, $\sigma_0 = 1$ and obtain up-to-scale reconstructions. Our calibration errors are between ± 3 gray levels.

3.3.2 Self-Supervision Losses

Formally, the network of Fig. 3.1 takes as input an image $I \in [0,1]^{w \times h \times 3}$, estimates a depth map $\widehat{d} \in (0, \infty)^{w \times h}$ and an albedo map $\widehat{\rho} \in [0,1]^{w \times h \times 2}$. It infers normals $\widehat{\mathbf{n}}$ from $\widehat{d}$, and uses $\widehat{d}$, $\widehat{\mathbf{n}}$, and ρ to render an image $\widehat{I} \in [0,1]^{w \times h \times 3}$ that should be as similar as possible to I. To train this network, we minimize a loss

$$\mathcal{L} = \mathcal{L}_p + \lambda_s \mathcal{L}_s + \lambda_{sp} \mathcal{L}_{sp} , \tag{3.4}$$

where λ_s and λ_{sp} are scalar weights and $\mathcal{L}_p$, $\mathcal{L}_s$, and $\mathcal{L}_{sp}$ are the loss terms described below.

$\mathcal{L}_p$ is a photometric loss and we take it to be the squared L_2 distance between the

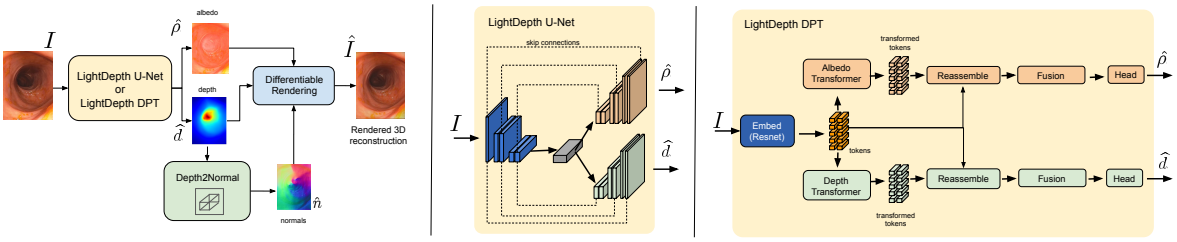


Figure 3.3: **Network Architecture. Left.** The input image is fed into a neural network that predicts albedo and depth values for each pixel. From the estimated depths, we compute the normals at each pixel surface using a kernel-based approach. Then, the depths, albedos, and normals are sent to a differentiable renderer that takes into account illumination decline and the endoscope's photometric model, and generates a synthetic image that should be as similar as possible to the original one. We also use specular reflections in saturated pixels to self-supervise normals. We investigated two different architectures: **Center.** LightDepth U-Net is based on a standard U-Net (Ronneberger et al., 2015) with two decoding branches. **Right.** LightDepth DPT is based on the DPT-Hybrid architecture (Ranftl et al., 2021), with a second decoder branch added for the albedo.

original image I and the rendered one $\hat{I}$. Note that because our rendering model is fully differentiable, we can perform end-to-end training.

$$\mathcal{L}_p = \sum_{i \in \Omega} (I_i - \widehat{I}_i)^2, \quad \text{where} \quad \widehat{I}_i = \mathcal{I}(i, \widehat{d}_i, \widehat{\rho}_i, g) \tag{3.5}$$

As in Godard et al. (2019), $\mathcal{L}_s$ is a regularization term that minimizes depth gradients except in areas of high color gradients, that may correspond to depth discontinuities. We write

$$\mathcal{L}_s = |\partial_x \widehat{d}| e^{-|\partial_x I|} + |\partial_y \widehat{d}| e^{-|\partial_y I|} \tag{3.6}$$

Finally, recall that we made a Lambertian assumption in Eq. 3.3, which prevents us to account properly for specular reflections and the overexposed pixels they produce. This is a potential source of error and fails to exploit the very useful information that specularities provide about normals. To remedy this, we introduce specular loss $\mathcal{L}_{sp}$. Given image location i, the corresponding direction $\mathbf{l}_i$ from the surface to the light source and the normal of a the surface $\widehat{\mathbf{n}}$, the law of reflection states that

$$\mathbf{s}_i = \mathbf{l}_i - 2\hat{\mathbf{n}}_i \left(\hat{\mathbf{n}}_i \cdot \mathbf{l}_i\right) \tag{3.7}$$

is the specularly reflected direction. Hence, we take our specular loss term to be

$$\mathcal{L}_{sp} = \sum_{i \in \Omega} \left(m_i \left(\mathbf{s}_i \cdot (-\mathbf{r}_i) - 1\right)\right)^2 , \tag{3.8}$$

$$m_i = \begin{cases} 1 & I_i > th \\ 0 & \text{otherwise} \end{cases}$$

which minimizes the discrepancy between the expected specular reflection $\mathbf{s}_i$ and the actual direction $(-\mathbf{r}_i)$ where the camera observes the reflection, resulting in pixel with high intensity $th = 0.98$.

Our method takes a single image as input, which makes 3D shape recovery solely from pixel colors an underconstrained problem. According to Eq. 3.3, a change in the brightness of a pixel can be due to changes in depth, albedo, camera exposure or surface normal. For example, if a given pixel is very bright, it can be because the pixel is close to the camera/light; the surface has a different albedo, resulting in more light being reflected; the surface normal is aligned to the light/camera, which increases the reflected light; the camera exposure and digital gain have been increased, which impacts brightness values in the whole image. Given the albedo at each surface point and the camera auto-gain, we could resolve these ambiguities. However, in medical endoscopy, true albedos are unknown, and auto-gain is not provided by the hardware manufacturer.

Albedo Constancy. We observe that endoscopy images exhibit a limited range of colors, with brighter tones being present in close areas and darker tones in deeper regions. Consequently, we hypothesize a significant correlation between albedo and the chromatic attributes, namely Hue and Saturation, in the HSV color space, as well as between depth and the Value Channel. In this way, we constrain the palette of colors that can be explained by the albedo decoder and we enhance the disentanglement between depth and albedo by setting $V = 100$ for all albedo values. Hence, to predict the albedo map $\widehat{\rho}$, our network predicts just two channels per pixel, for Hue and Saturation, and assumes Value to be one to convert to the RGB space, in which the loss is formulated.

3.3.3 Network Architecture

Our network outputs depth and albedo maps. In Fig. 3.3, we provide a more detailed depiction of our encoder-decoder architecture. We have tested two different versions. The first one is a U-net with two decoders and skip connections, with a ResNet18 serving as the backbone. Our decoders design is inspired by Godard et al. (2019). The second one relies on visual transformers for depth estimation (Ranftl et al., 2021). As a backbone, we use a Resnet-50 (DPT-Hybrid) and two decoders that reassemble the tokens and apply attention heads. Further details regarding these architectures can be found in section 3.8.

In both versions, to compute the normals at any given pixel, we use a convolution kernel with six-neighborhood (N, NE, E, S, SW, and W) in the depth map. We define

six triangles using the central pixel as reference, with each triangle having its own normal. The normal of the central pixel is computed as the average of the normals of the triangles weighted by their area. The use of six neighbors lets us reuse triangles during the convolution pass to speed up computation.

3.4 Results

3.4.1 Datasets

We evaluated LightDepth and relevant baselines on three endoscopy datasets: An in-house *synthetic colon*, *C3VD* (Bobrow et al., 2023), and *EndoMapper* (Azagra et al., 2023). With these, we can show quantitative and qualitative results with several levels of realism.

Synthetic Colon. We simulate a real Olympus CF-H190L endoscope consisting on a fish-eye camera and a spot light source, both calibrated as in Azagra et al. (2023). This is in contrast to other synthetic datasets that simulate arbitrary camera and illumination configurations, typically pinhole cameras with no or arbitrary distortion and ideal light sources with no radial falloff. (Rau et al., 2019; Zhang et al., 2020; Ozyoruk et al., 2021; Rau et al., 2023). We rendered the images using ray-casting techniques, in which the colon's geometry and albedo are defined by a triangle mesh obtained from a CT scan of a real colon (İncetan et al., 2021). We ignore global illumination effects and assume Lambertianity, so there are no specular reflections. The influence of these two effects will be assessed in the two other datasets. Our synthetic data is hence composed by 1620 fish-eye RGB frames annotated with per-pixel albedo, depth and normals. We split it into 1168 images for training and 452 images for test. Example frames can be found in section 3.7.

C3VD (Bobrow et al., 2023) contains real images recorded in a phantom with ground-truth depth. The images have been captured by a real Olympus CF-HQ190L endoscope in a phantom silicone model of a human colon. The data is annotated with ground-truth depth and normals by applying 2D-3D registration of the 3D phantom models. The authors claim that the silicone material is opaque, hence we can assume that the only light source available is in the endoscope. Finally, it includes a geometrical calibration based on Scaramuzza et al. (2006). C3VD provides a good compromise between realism (real endoscope, global illumination effects and specular highlights) and ground-truth labels for quantitative evaluation. Of the 10,088 images available, we use 7,200 for training and 2,888 for testing. In section 3.8, we provide the sections of the phantom used for testing and training.

EndoMapper (Azagra et al., 2023) provides the most challenging data, as it con-

				Depth [mm]									Normals [°]
Dataset	Architecture	Backbone	Supervision	MAE ↓	MedAE ↓	RMSE ↓	$\text{RMSE}_{\log}$ ↓	Abs_{Rel} ↓	Sq_{Rel} ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑	MAE ↓
Synthetic	U-Net	ResNet18	Depth GT	**4.37**	2.99	**6.38**	**0.1251**	0.0965	**0.0008**	0.9057	**0.9931**	**0.9997**	25.1
	LightDepth U-Net	ResNet18	Light	4.76	**2.47**	8.60	0.1375	**0.0903**	0.0011	**0.9180**	0.9820	0.9935	**15.2**
C3VD	U-Net	ResNet18	Depth GT	4.15	3.29	5.52	0.1139	0.0902	0.0007	0.9172	0.9943	**0.9994**	26.5
	DPT-Hybrid (2021)	ResNet50	Depth GT	**3.22**	2.77	**4.10**	**0.0860**	**0.0699**	**0.0004**	**0.9640**	0.9865	0.9913	**15.1**
	Monodepth2 (2019)	ResNet50	Multi-View	14.27	9.59	18.64	0.3921	0.2971	0.0070	0.4897	0.7313	0.8611	43.6
	CADepth (2021)	ResNet18	Multi-View	52.35	17.04	87.43	0.9144	1.1916	0.2650	0.3664	0.5653	0.6679	67.2
	XDCycleGAN (2020)	ResNet	Cycle	17.16	11.91	22.43	0.4953	0.3616	0.0105	0.4291	0.6615	0.7910	64.4
	LightDepth U-Net	ResNet18	Light	4.37	2.92	6.31	0.1183	0.0856	0.0007	0.9315	0.9934	**0.9994**	24.0
	LightDepth DPT	ResNet50	Light	3.94	2.67	5.60	0.1080	0.08046	0.0006	0.9476	0.9965	**0.9994**	21.3
	LightDepth U-Net	ResNet18	Light (TTR)	3.72	2.59	5.43	0.1060	0.0770	0.0005	0.9505	**0.9971**	**0.9994**	23.5
	LightDepth DPT	ResNet50	Light (TTR)	3.70	**2.58**	5.27	0.1073	0.0780	0.0005	0.9525	0.9961	0.9992	22.5

Table 3.1: Depth and normal metrics for several architectures and supervision modes.
Best results per dataset are bolfaced, second best underlined.

Dataset	Architecture	Supervision	SSIM ↑	MAE ↓
Synthetic	LightDepth U-Net	Light	0.9901	0.0192
C3VD	LightDepth U-Net	Light	0.9765	0.0657
	LightDepth DPT	Light	0.8873	0.0599
	LightDepth U-Net	Light (TTR)	**0.9811**	**0.0276**
	LightDepth DPT	Light (TTR)	0.8977	0.0329

Table 3.2: SSIM and MAE for rendered images in C3VD.
Test-time refinement (TTR) gives a substantial improvement.

tains real colonoscopy and gastroscopy procedures inside the human body, performed by endoscopists on a day-to-day basis. Here we find real textures such as veins, blood and dirt, and other effects such as blur, water and frames very close or even hitting the mucosa. Foam and bubbles are indeed very common in endoscopy images and are usually ignored. LightDepth is capable of disentangling these as part of the albedo and not of the depth. Before processing the dataset, we perform a manual inspection of the selected sequences and we eliminate occluded and excessively blurred frames.

Finally, we train in three procedures, consisting of two colonoscopies and one gastroscopy. There are a total of 24,444, 23,456 and 3,032 frames, respectively. Details of the sequences and frames we use can be found in section 3.8.

3.4.2 Metrics, Baselines, and Training Details

We report results using a median-based scale alignment for all methods, even those supervised with real-scale depth, for fairness. In our experiments, we compare against models that use depth supervision and multi-view self-supervision. For depth supervision, we use two different architectures, U-Net with L1 loss as a representative of convolutional architectures and DPT-Hybrid (Ranftl et al., 2021) as a state-of-the-art representative of transformer-based models, learning inverse depth with an scale invariant loss.

For a fair comparison, we also evaluate our LightDepth using the same U-Net and DPT architectures. The U-Net is pre-trained on ImageNet dataset (Deng et al., 2009). For DPT, we initialize with the author-provided weights for encoder and depth decoder. The albedo decoder is trained from scratch. During training, we select a smoothing weight $\lambda_s = 0.1$ in Eq. 3.4 and a learning rate of 10^{-4} for the Adam optimizer. In the synthetic dataset, we trained our network with $\lambda_{sp} = 0$, as synthetic dataset has no specular reflections. In C3VD and EndoMapper, we use $\lambda_{sp} = 1$.

Test-Time Refinement (TTR). As our LightDepth enables single-view self-supervision, we can continuously refine the depth predictions online, obtaining much more accurate reconstructions. In the results denoted as "(TTR)", we perform online test-time refinement for each test image separately during $N = 20$ optimization steps, using the loss $\mathcal{L}$ in Equation 3.4, as in training time. To mitigate the risk of catastrophic forgetting, we load again the original model trained in the train split after TTR for each image.

Note in Table 3.1 how TTR improves significantly the metrics with respect to LightDepth without TTR for U-Net and DPT architectures. Remarkably, observe how TTR even outperforms the metrics achieved by Depth GT supervision. Figure 3.4 shows the improvement given by TTR in the network prediction of depth, normals and

albedo and overall in the 3D reconstruction. Inference time is $\sim$ 5ms for LightDepth U-Net and $\sim$ 22 ms for LightDepth DPT on a NVIDIA GeForce RTX 3090. We can do TTR in $\sim$ 90 ms per optimization step in U-Net and $\sim$ 190 ms in DPT.

3.4.3 Quantitative Results on Synthetic and Phantom

Synthetic colon. The first two rows in Table 3.1 report depth and normal metrics for a U-Net supervised with Depth GT, and our self-supervised LightDepth U-Net architecture. Observe that the metrics are similar. This is notable, as self-supervision is consistently reported in the literature to underperform with respect to depth supervision, and suggests that illumination decline provides a very strong self-supervisory signal in endoscopies, which our experiments in the other two datasets confirm.

Furthermore, light self-supervision outperforms Depth GT supervision in MedAE and $\delta < 1.25$, which means that most of the error distribution is lower for light self-supervision and only a small fraction of large errors are better with depth supervision. We observed that it is in far and dark areas where light self-supervision is weaker and this produces a higher depth MAE and RMSE. Observe the significantly lower error in normal with our light self-supervision, due to the lower errors in most pixels.

C3VD Phantom. We report depth and normal metrics on the real phantom images of the C3VD dataset in Table 3.1. Our self-supervised architectures LightDepth U-Net and LightDepth DPT with TTR outperform supervision with Depth GT in MedAE, while the rest of the metrics are very close. As in the case of the synthetic dataset, this is a remarkable result because self-supervised architectures typically lag behind supervised ones in single view depth estimation. The fact that LightDepth MedAE is better and RMSE is worse suggests that our errors are better in most of the distribution, and there are a few regions with large errors where Depth GT supervision is able to offer an advantage. Table 3.2 details metrics on the quality of the rendered image, which suggest the strength of the self-supervision signal. Observe the improvement of this metrics for the TTR case.

In Table 3.1, observe that the multi-view self-supervised baselines, Monodepth2 (Godard et al., 2019) and CADepth (Yan et al., 2021), have a poor performance in our data, worse in comparison than results in other datasets. This could be due to the weak textures and changing lighting in the colonoscopy images, resulting in noisy estimations for relative motion and uninformative photometric residuals. Being single-image, our approach is impervious to such difficulties.

Domain shift. As synthetic-to-real is common in endoscopies to address the lack of ground-truth depth for supervision, we also evaluated XDCycleGAN (Mathew et al.,

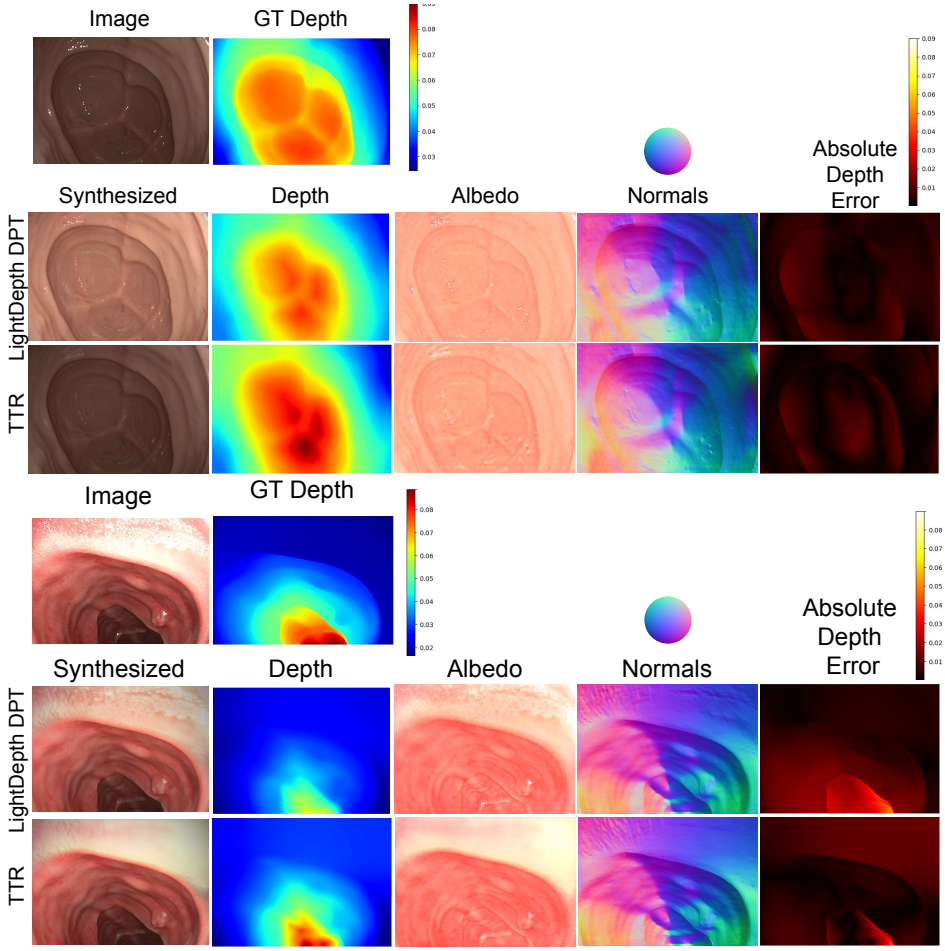


Figure 3.4: LightDepth and LightDepth TTR on C3VD. Our light decline captures the correct shape of the cecum in the first image and the shape of the polyp in the second. Note how the estimates of normals and albedo are similar before and after TTR. By optimising depth by reducing illumination, LightDepth achieves a darker appearance and improvements in depth estimation.

Dataset			Depth [mm]			Normals [°]
Train	Test	Supervision	MAE ↓	MedAE ↓	RMSE↓	MAE ↓
Synt.	Synt.	Depth GT	4.37	2.99	6.38	25.1
		Light	4.76	2.47	8.60	15.2
Synt.	C3VD	Depth GT	9.44	5.79	12.83	73.7
		Light	5.09	3.51	7.14	27.7
		Depth GT (TTR)	4.96	3.14	7.11	25.4
		Light (TTR)	<u>3.80</u>	**2.51**	5.54	<u>23.6</u>
C3VD	C3VD	Depth GT	4.15	3.29	<u>5.52</u>	26.5
		Light	4.37	2.92	6.31	24.0
		Light (TTR)	**3.72**	<u>2.59</u>	**5.43**	**23.5**

Table 3.3: Synthetic-to-real domain shift. Best results in C3VD test set are boldfaced, second best are underlined. Note the domain shift effect between Synt. and C3VD test data in the bigger errors, and how TTR removes the domain shift effect completely. Notably, our LightDepth TTR delivers similar errors than the models without domain shift, trained in C3VD.

Method	MAE [°]
U-Net	16.24
TFtN (Fan et al., 2021)	3.89
Open3D (Zhou et al., 2018a)	1.67
In-house	**1.32**

Table 3.4: Normal accuracy for baseline methods and LightDepth.

2020) as a baseline. Note that the domain shift is still affecting the results. Our single-view LightDepth self-supervision enables training in the target domain, and hence removes completely the domain shift, achangedchieving significantly lower errors.

Table 3.3 elaborates further on domain shift by showing depth and normals metrics for a U-Net architecture in these cases. Specifically, we trained a U-Net model with Depth GT supervision and light self-supervision in our synthetic dataset and evaluated their performance in the synthetic and C3VD test sets. Observe how the domain shift affects all metrics significantly. Interestingly, the model trained with light self-supervision and without TTR generalizes significantly better to the C3VD data, as our LightDepth self-supervised model is closer to the physical phenomena than Depth GT supervision. Again, note that single-view self-supervision removes completely the domain shift effect, as models can be trained directly in the target domain. Very remarkably, the performance of our models with domain shift after TTR matches the performance of the models without domain shift.

Normals from Depth. The literature details different manners to obtain surface normals from a depth map, e.g., (Boulch and Marlet, 2016; Fan et al., 2021). Table 3.4 shows a MAE analysis of the most promising ones in C3VD. Specifically, we eval-

Loss	Depth [mm] MAE ↓	MedAE ↓	RMSE↓	Color MAE ↓	Normals [°] MAE ↓
$\mathcal{L}_p$	6.05	3.93	8.79	0.0637	35.5
$\mathcal{L}_p + \mathcal{L}_s$	4.95	3.04	7.23	0.0690	24.6
$\mathcal{L}_p + \mathcal{L}_s + \mathcal{L}_{sp}$	4.37	2.92	6.31	0.0657	24.0

Table 3.5: Ablation study of the losses with LightDepth U-Net in C3VD dataset. Observe the improvement given by each term.

uate four methods: a U-net trained to regress normals from depth, the recent TFtN method (Fan et al., 2021), the implementation in Open3D (Zhou et al., 2018b) that computes normals from a k-nearest neighbourhood in the point cloud, and an in-house method that uses six-neighbourhood in the image. Our analysis shows that an analytic average in a neighbourhood is significantly better than a U-Net and TFtN, and our in-house method that considers a neighbourhood in the image is slightly better, so this last one was our choice.

Ablation Study on the Loss. In Table 3.5, we ablate the terms of our loss function. The smoothness prior ($\mathcal{L}_s$ term) is remarkably beneficial for both depth and normal prediction. When we do not take advantage of the information of the specular reflections (no $\mathcal{L}_{sp}$ term), we obtain worse results. Adding this new loss term, we see how all the depth and normal metrics improve, especially in the median error, which outperforms the supervised and now matches that obtained in the simulation experiment. Still, the depth MAE and RMSE are slightly higher than those of the baseline due to the far spurious points.

3.4.4 Qualitative Results in Real Endoscopy

We now turn to real images of a human colon from the EndoMapper dataset and present qualitative results in Figure 3.5. Additional ones can be found in section 3.7. Some details are recovered very accurately, such as the normal maps showing clearly the tubular shape; the depth maps reflecting the discontinuities in the haustra; the albedos capturing the blood vessels, in particular in the 5^{th} column; and the bubbles and fluids colors in the 6^{th} and 7^{th} columns, which make the 3d reconstruction of these bubbles and fluids very plausible.

Unfortunately, there is no ground-truth data available for this dataset, which prevent us from presenting quantitative results, and we do not know of any other dataset with real colonoscopy images that includes ground-truth data. Nevertheless, visual inspection of our results hints that the strengths of our techniques demonstrated quantitatively in Section 3.4.3 will carry over on truly realistic scenarios like this one.

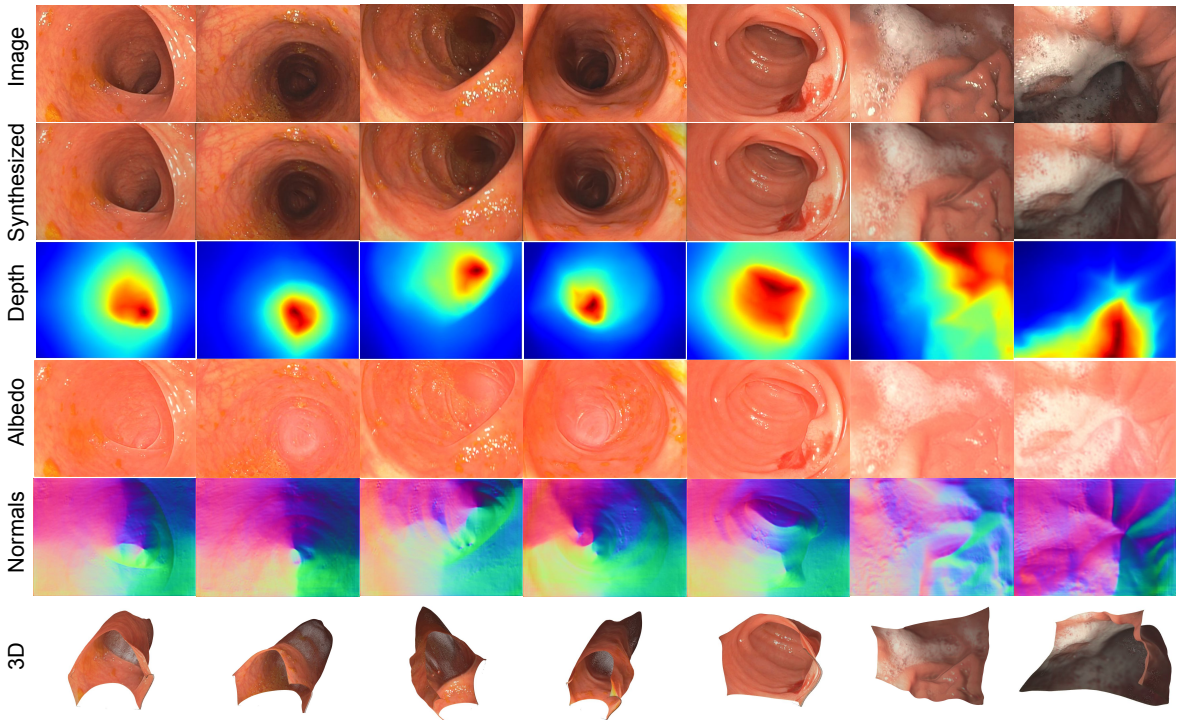


Figure 3.5: Qualitative results on EndoMapper with LightDepth DPT. Columns 1–5 are real colonoscopy images, and columns 6–7 are real gastroscopy images. In colonoscopies, observe that the normals exhibit a tubular shape specific of the colon. The albedo prediction captures disruptions such as veins, blood, dirt, foam and specularites. Note the influence of light decline in the image and the correlation with the estimated depths.

3.5 Limitations and Discussion

As mentioned in Section 3.3.2, our depth predictions are up-to-scale. Even if the camera auto-gain was available, the albedo scale may be challenging to learn, so estimating the real scale is not straightforward. In any case, other methods such as multi-view self-supervision or synthetic-to-real cannot guarantee an accurate estimation of the scale either. We assume that Lambertian reflectance is prevalent in most tissues, and for areas where this does not hold, we use a basic model to capture specularities. Further research could focus on the application of more sophisticated photometric models that cover specularities, e.g., the Phong model.

Thanks to our priors on albedo and depth, we successfully disentangle both factors in our experiments. However, our $V = 100$ prior might not hold in areas of clotted blood or with very dark albedos, e.g., because of a disease. These priors might need to be tuned in new application domains for enhanced performance. Finally, although we demonstrate this technology in the context of endoscopy, its principles are applicable in any setup in which the only light source is close to the target surface and rigidly attached to the camera. In other words, our LightDepth has the potential to open research avenues in many other domains.

3.6 Conclusions

In this work, we have proposed, for the first time, a single-view self-supervision method for depth learning, which we denote LightDepth, that exploits and is limited to the case of a single spotlight source co-located with a monocular camera, a case that includes, among others, the relevant application of medical endoscopy. As our main contribution, we developed the specific self-supervised learning setup that models the quadratic light decline and enables self-supervised learning. We have implemented two different architectures, a first one based on convolutions and a second one based on transformers, and evaluated their performance against ground-truth supervision, multi-view self-supervision, and domain transfer approaches. Our results show that LightDepth outperforms multi-view self-supervision and synthetic-to-real transfer and matches the performance of fully supervised approaches. Not only that, its training and test-time refinement setup is significantly simpler: LightDepth only requires a reasonable endoscope calibration and does *not* require camera motion estimation nor ground-truth labels nor realistic simulations, all of them challenging in endoscopies. This unlocks, from a practical point of view, relevant potential applications in the medical domain.

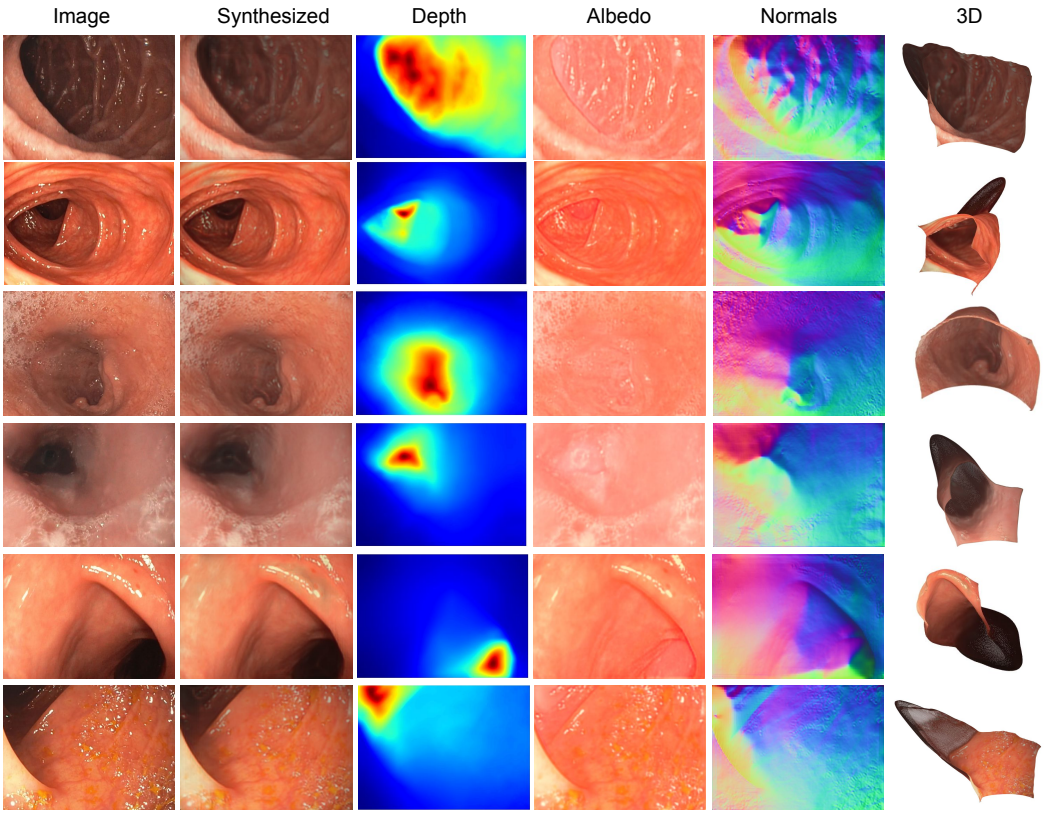


Figure 3.6: Additional examples of LightDepthDPT in real procedures

3.7 Additional results

We present additional quantitative and qualitative results. Figure 3.6 shows additional qualitative results of LightDepthDPT TTR in real colonoscopy and gastroscopy procedures. Figure 3.7 shows quantitative results of LightDepth U-Net in the transverse and cecum sections of the C3VD.

Finally, in Figure 3.8 we show examples of LightDepth U-Net in our in-house synthetic dataset. The predicted depth and normals capture the shape of the colon sections, as shown in the 3D reconstruction. The albedo map appears brighter as we fix Value Channel to $V = 100$. Our method recovers the different albedo of mucosa and blood vessels.

3.8 Implementation details

3.8.1 Network architectures

LightDepth U-Net. We use a U-Net architecture with skip connections and two decoders. Our encoder is a ResNet18 (He et al., 2016) initialized with the weights from ImageNet (Deng et al., 2009). Regarding the decoders, our albedo decoder uses sigmoid activation and our depth decoder ELU+1 after the last convolution.

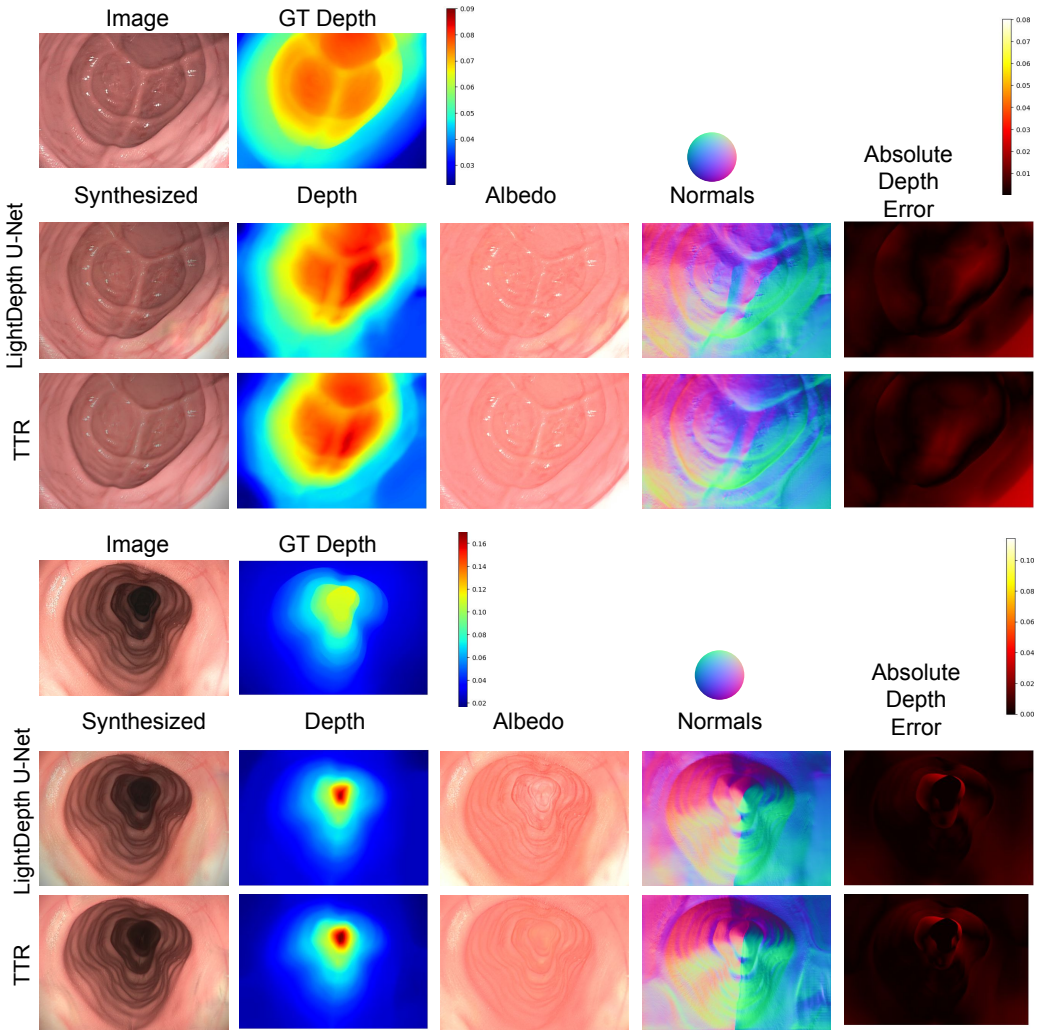


Figure 3.7: Additional examples of LightDepth U-Net in C3VD

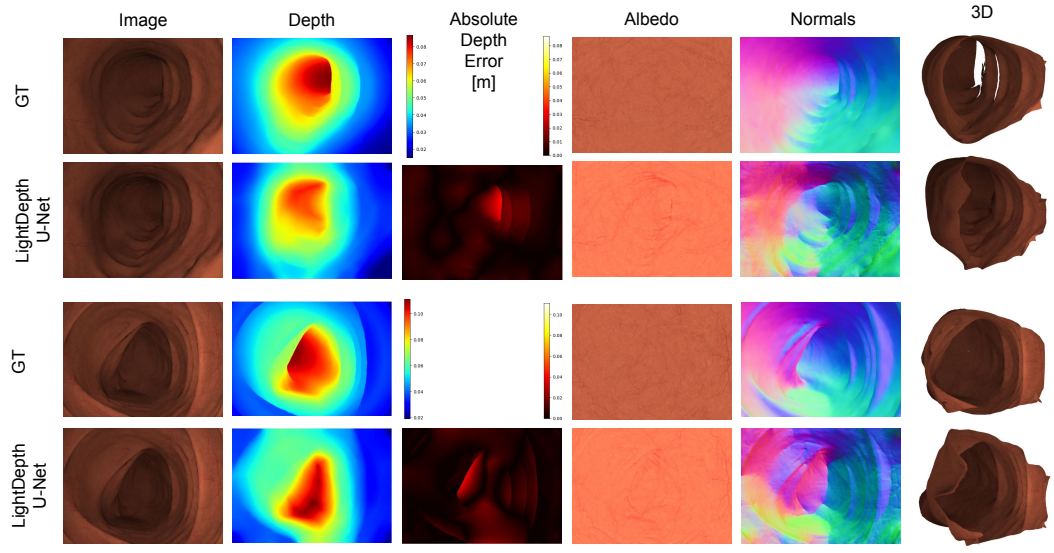


Figure 3.8: Qualitative examples of LightDepth in Synthetic dataset

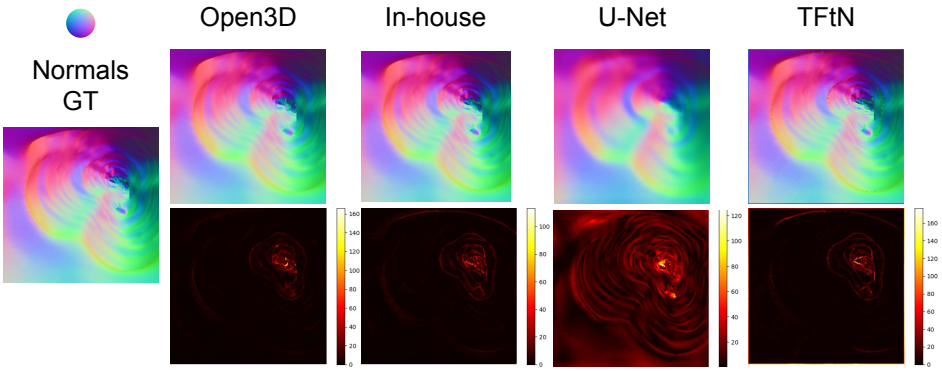


Figure 3.9: Quantitative results of surface normals estimation from a depth map

LightDepth DPT. We extend LightDepth DPT (Ranftl et al., 2021) adding a branch for the prediction of albedo decoder. For the depth estimation, we initialize the encoder and depth decoder with DPT Hybrid weights. For albedo estimation, we train the albedo decoder from scratch. In our pipeline, we use the half of resolution than the original images for training, up sampling the outputs with bilinear interpolation.

3.8.2 Datasets

Table 3.6 shows which sections of the C3VD were used for training / testing. We split into sections to ensure a fair comparison along the dataset. Regarding real endoscopy images, we use with the sequence 051, 009 and 058 of the EndoMapper dataset.

3.8.3 Normals from Depth

Figure 3.9 shows examples of Open3D (Zhou et al., 2018b), in-house, U-Net and TFtN (Fan et al., 2021) used in the analysis. Figure 3.10 presents the head for the albedo decoder that includes a sigmoid activation function.

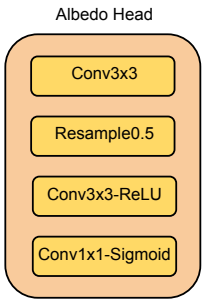


Figure 3.10: Albedo estimation head in LightDepth

Model	Texture	Video	Frames	Stage
Cecum	1	b	765	Train
Cecum	2	b	1120	Train
Cecum	2	c	595	Train
Cecum	4	a	465	Train
Cecum	4	b	425	Train
Sigmoid Colon	1	a	800	Train
Sigmoid Colon	2	a	513	Train
Sigmoid Colon	3	b	536	Train
Transcending Colon	1	a	61	Train
Transcending Colon	1	b	700	Train
Transcending Colon	2	b	102	Train
Transcending Colon	4	b	595	Train
Descending Colon	4	a	74	Train
Cecum	1	a	275	Test
Cecum	2	a	370	Test
Cecum	3	a	730	Test
Descending Colon	4	a	74	Test
Sigmoid Colon	3	a	610	Test
Transcending Colon	2	a	194	Test
Transcending Colon	3	a	250	Test
Transcending Colon	4	a	382	Test

Table 3.6: Dataset split for C3VD

Chapter 4

Neural Surface Reconstruction from Illumination Decline

4.1 Introduction

Colorectal cancer (CRC) is the third most commonly diagnosed cancer and is the second most common cause of cancer death (Sung et al., 2021). Early detection is crucial for a good prognosis. Despite the existence of other techniques, such as virtual colonoscopy (VC), optical colonoscopy (OC) remains the gold standard for colonoscopy screening and the removal of precursor lesions. Unfortunately, we do not yet have the ability to reconstruct densely the 3D shape of large sections of the colon. This would usher exciting new developments, such as post-intervention diagnosis, measuring polyps and stenosis, and automatically evaluating exploration thoroughness in terms of the surface percentage that has been observed.

This is the problem we address here. It has been shown that the colon 3D shape can be estimated from single images acquired during human colonoscopies (Batlle et al., 2022). However, to model large sections of it while increasing the reconstruction accuracy, multiple images must be used. As most endoscopes contain a single camera, the natural way to do this is to use video sequences acquired by these cameras in the manner of structure-from-motion algorithms. An important first step in that direction is to register the images from the sequences. This can now be done reliably using either batch (Schönberger and Frahm, 2016) or SLAM techniques (Gómez-Rodríguez et al., 2021). Unfortunately, this solves only half the problem because these techniques provide very sparse reconstructions and going from there to dense ones remains an open problem. And occlusions, specularities, varying albedos, and specificities of endoscopic lighting make it a challenging one. To overcome these difficulties, we rely on two properties of endoscopic images:

- Endoluminal cavities such as the gastrointestinal tract, and in particular the

human colon, are watertight surfaces. To account for this, we represent its surface in terms of a signed distance function (SDF), which by its very nature presents continuous watertight surfaces.

- In endoscopy the light source is co-located with the camera. It illuminates a dark scene and is always close to the surface. As a result, the irradiance decreases rapidly with distance t from camera to surface; more specifically it is a function of $1/t^2$. In other words, there is a strong correlation between light and depth, which remains unexploited to date.

To take advantage of these specificities, we build on the success of Neural implicit Surfaces (Wang et al., 2021b, NeuS) that have been shown to be highly effective at deriving surface 3D models from sets of registered images. As the Neural Radiance Fields (Mildenhall et al., 2021, NeRF) that inspired them, they were designed to operate on regular images taken around a scene, sampling fairly regularly the set of possible viewing directions. Furthermore, the lighting is assumed to be static and distant so that the brightness of a pixel and its distance to the camera are unrelated. Unfortunately, none of these conditions hold in endoscopies. The camera is inside a cavity (in the colon, a roughly cylindrical tunnel) that limits viewing directions. The light source is co-located with the camera and close to the surface, which results in a strong correlation between pixel brightness and distance to the camera. In this chapter, we show that, far from being a handicap, this correlation is key for neural network self-supervision.

NeuS training selects a pixel from an image and samples points along its projecting ray. However, the network is agnostic to the sampling distance. In LightNeuS, we explicitly feed to the renderer the distance of each one of these sampled points to the light source, as shown in Figure 4.1. Hence, the renderer can exploit the inverse-square illumination decline. We also introduce and calibrate a photometric model for the endoscope light and camera, so that the inverse square law discussed above actually holds. Together, these two changes make the minimization problem better posed and the automatic depth estimation more reliable.

Our results show that exploiting the illumination is key to unlocking implicit neural surface reconstruction in endoscopy. It delivers accuracies in the range of 3 mm, whereas an unmodified NeuS is either 5 times less accurate or even fails to reconstruct any surface at all. Earlier methods in Batlle et al. (2022) have reported similar accuracies but only on very few synthetic images and on short sections of the colon. By contrast, we can handle much longer ones and provide a broad evaluation in a real dataset (C3VD) over multiple sequences. This makes us the first to show accurate results of extended 3D watertight surfaces from monocular endoscopy images.

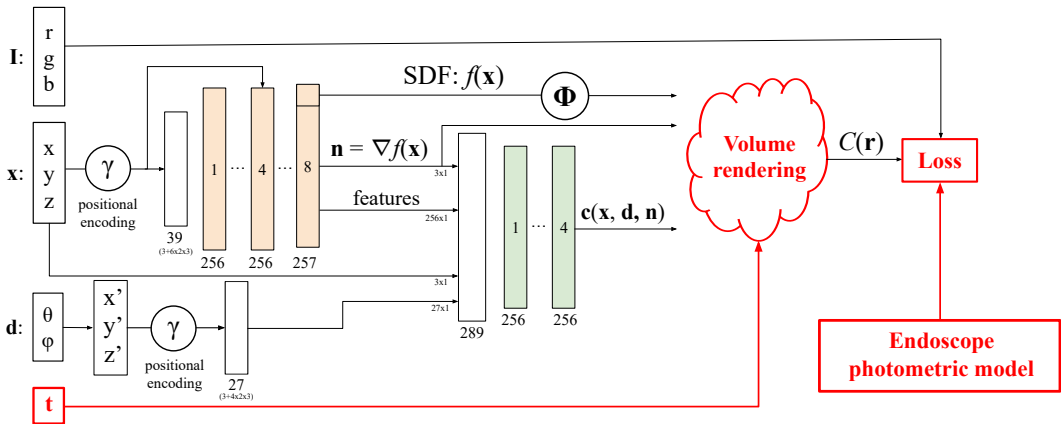


Figure 4.1: **From NeuS to LightNeuS**. The original NeuS architecture is depicted by the black arrows. In LightNeuS, when training the network with a sampled point, we provide the sampling distance t to the renderer, that takes into account illumination decline. We also incorporate a calibrated photometric endoscope model that is used to correctly compute the photometric loss. The changes are shown in red.

4.2 Related work

3D Reconstruction from Endoscopic Images. It can help with the effective localization of lesions, such as polyps and adenomas, by providing a complete representation of the observed surface. Unfortunately, many state-of the-art SLAM techniques based on feature matching (Campos et al., 2021) or direct methods (Engel et al., 2014, 2017) are impractical for dense endoscopic reconstruction due to the lack of texture and the inconsistent lighting that moves along with the camera. Nevertheless, sparse reconstructions by classical Structure-from-Motion (SfM) algorithms can be good starting points for refinement and densification based on Shape-from-Shading (SfS) (Tokgozoglu et al., 2012; Zhao et al., 2016). However, classical multi-view and SfS methods require strong suboptimal priors on colon surface shape and reflectance.

In monocular dense reconstructions, it is common practice to encode shape priors in terms of smooth rigid surfaces (Newcombe et al., 2011; Schönberger et al., 2016; Mahmoud et al., 2018). Recently, Sengupta and Bartoli (2021) proposes a tubular topology prior for NRSfM aimed to process endoluminal cavities where these tubular shapes are prevalent. In contrast, for the same environments, we propose the watertight prior coded by implicit SDF representations.

Recent methods for dense reconstruction rely on neural networks to predict per-pixel depth in the 2D space of each image and fuse the depth maps by using multi-view stereo (MVS) (Bae et al., 2020) or a SLAM pipeline (Ma et al., 2019, 2021). However, holes in the reconstruction appear due to failures in triangulation and inaccurate depth estimation or in areas not observed in any image. Wang et al. (2022) show the potential

of neural rendering in reconstruction from medical images, although they use a binocular static camera with fixed light source, which is not feasible in endoluminal endoscopy. Unfortunately, most of the previous 3D methods do not provide code (Mahmoud et al., 2018; Sengupta and Bartoli, 2021), are not evaluated in biomedical settings (Newcombe et al., 2011; Schönberger et al., 2016), or do not report reconstruction accuracy (Ma et al., 2019, 2021).

Neural Radiance Fields (NeRFs) were first proposed to reconstruct novel views of non-Lambertian objects (Mildenhall et al., 2021). This method provides an *implicit neural representation* of a scene in terms of local densities and associated colors. In effect, the scene representation is stored in the weights of a neural network, usually a multilayer perceptron (MLP), that learns its shape and reflectance for any coordinate and viewing direction. NeRFs use volume rendering (Kajiya and Von Herzen, 1984), based on ray-tracing from multiple camera positions. The volume density $\sigma(\mathbf{x})$ can be interpreted as the differential probability of a ray terminating at an infinitesimal particle at location $\mathbf{x}$. The expected color $C(\mathbf{r})$ of the pixel with camera ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ is the integration of the radiance emitted by the field at every traveled distance t from near to far bounds t_n and t_f, such that

$$C(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\, \sigma(\mathbf{r}(t))\, \mathbf{c}(\mathbf{r}(t), \mathbf{d})\, \mathrm{d}t \quad \text{where } T(t) = \exp\left(- \int_{t_n}^{t} \sigma(\mathbf{r}(s))\, \mathrm{d}s \right) \tag{4.1}$$

where $\mathbf{c}$ stands for the color. The function T denotes the accumulated transmittance along the ray from t_n to t, that is the probability that the ray travels from t_n to t without hitting any other particle. The authors propose two MLPs to estimate the volume density function $\sigma : \mathbf{x} \to [0, 1]$ and the directional emitted color function $\mathbf{c} : (\mathbf{x}, \mathbf{d}) \to [0, 1]^3$, so the density of a point does not depend on the viewing direction $\mathbf{d}$, but the color does. This allows them to model non-Lambertian reflectance. In addition, they propose a positional encoding for location $\mathbf{x}$ and direction $\mathbf{d}$, which allows high-frequency details in the reconstruction.

Neural Implicit Surfaces (NeuS) were introduced in Wang et al. (2021b) to improve the quality of NeRF representation modelling watertight surfaces. For that, the volume density σ is computed so as to be maximal at the zero-crossings of a signed distance function (SDF) f:

$$\sigma(\mathbf{r}(t)) = \max\left(\frac{\Phi_s'(f(\mathbf{r}(t)))}{\Phi_s(f(\mathbf{r}(t)))}, 0 \right) \quad \text{where } \Phi_s(x) = \frac{1}{1 + e^{-sx}} \tag{4.2}$$

The SDF formulation makes it possible to estimate the surface normal as $\mathbf{n} = \nabla f(\mathbf{x})$. The reflectance of a material is usually determined as a function of the incoming and outgoing light directions with respect to the surface normal. Therefore, the

normal is added as an input to the MLP that estimates color $\mathbf{c} : (\mathbf{x}, \mathbf{d}, \mathbf{n})$, as shown in Figure 4.1.

4.3 LightNeuS

In this section, we present the key contributions that make *LightNeuS* a neural implicit reconstruction method suitable for endoscopy in endoluminal cavities. In this context, the light source is located next to the camera and moves with it. Furthermore, it is close to the surfaces to be modeled. As a result, for any surface point $\mathbf{x} = \mathbf{o} + t\mathbf{d}$, the irradiance decreases with the square of the distance to the camera t. Hence, we can write the color of the corresponding pixel as in Batlle et al. (2022):

$$\mathcal{I}(\mathbf{x}) = \left(\frac{L_e}{t^2} \, \mathrm{BRDF}(\mathbf{x}, \mathbf{d}) \, \cos(\theta) \, g \right)^{1/\gamma} \tag{4.3}$$

where L_e is the radiance emitted by the light source to the surface point, that was modeled and calibrated in the EndoMapper dataset (Azagra et al., 2023) according to the SLS model from Modrzejewski et al. (2020). The bidirectional reflectance distribution function (BRDF) determines how much light is reflected to the camera, and the cosine term $\cos(\theta) = -\mathbf{d} \cdot \mathbf{n}$ weights the incoming radiance with respect to the surface normal $\mathbf{n}$. Equation (4.3) also takes into account the camera gain g and gamma correction γ.

4.3.1 Using Illumination Decline as a Depth Cue

The NeuS formulation of section 4.2 assumes distant and fixed lighting. However, in endoscopy inverse-square light decline is significant, as quantified in Equation 4.3. Accounting for this is done by modifying the original NeuS formulation as follows. Figure 4.1 depicts the original NeuS network in black. It uses a SDF network—shown in orange—to estimate a view-independent geometry and only the final RGB color depends on the viewing direction $\mathbf{d}$. It is estimated by the network shown in green. Thus, this second network $\mathbf{c}(\mathbf{x}, \mathbf{d}, \mathbf{n})$ may learn to model non-Lambertian $\mathrm{BRDF}(\mathbf{x}, \mathbf{d})$, including specular highlights, and the cosine term of Equation 4.3. However, if the distance t from the light to the point $\mathbf{x}$ is not provided to the color network, the $1/t^2$ dependency cannot be learned, and surface reconstruction will fail. Our key insight is to explicitly supply this distance as input to the volume rendering algorithm, as shown in red in Figure 4.1 and reformulate Equation 4.1 as

$$C(\mathbf{r}) = \int_{t_n}^{t_f} T(t) \, \sigma(\mathbf{r}(t)) \, \frac{\mathbf{c}(\mathbf{r}(t), \mathbf{d}, \mathbf{n})}{t^2} \, \mathrm{d}t \tag{4.4}$$

This conceptually simple change, using illumination decline while training, unlocks all the power of neural surface reconstruction in endoscopy.

4.3.2 Endoscope Photometric Model

Apart from illumination decline, there are several significant differences between the images captured by endoscopes and those conventionally used to train NeRFs and NeuS: fish-eye lenses, strong vignetting, uneven scene illumination, and post-processing.

Endoscopes use fisheye lenses to cover a wide field of view, usually close to 170 degrees. These lenses produce strong deformations, making it unwise to use the standard pinhole camera model. Instead, specific models (Scaramuzza et al., 2006; Kannala and Brandt, 2006) must be used. Hence, we also modified the original NeuS implementation to support these models.

The light sources of endoscopes behave like spotlights. In other words, they do not emit with the same intensity in all directions, so L_e in Equation 4.3 is not constant for all image pixels. This effect is similar to the vignetting effect caused by conventional lenses, that is aggravated in fisheye lenses. Fortunately, they can be accurately calibrated (Modrzejewski et al., 2020; Azagra et al., 2023) and compensated for.

The post-processing software of medical endoscopes is designed to always display well-exposed images, so that physicians can see details correctly. An adaptive gain factor g is applied by the endoscope's internal logic and gamma correction is also used to adapt to non-linear human vision, achieving better contrast perception in mid tones and dark areas. Endoscope manufacturers know the post-processing logic of their devices, but this information is proprietary and not available to users. Again, gamma correction can be calibrated assuming it is constant (Batlle et al., 2022), and the gain change between successive images can be estimated, for example, by sparse feature matching.

All these factors must be taken into account during network training. Thus, our photometric loss is computed using a normalized image:

$$I' = \left(\frac{I^\gamma}{L_e g} \right)^{1/\gamma} \tag{4.5}$$

4.4 Experiments

We validate our method on the C3VD dataset (Bobrow et al., 2023), which covers all different sections of the colon anatomy in 22 video sequences. This dataset contains sequences recorded with a medical video colonoscope, Olympus Evis Exera III CF-HQ190L. The images were recorded inside a *phantom*, a model of a human colon made of silicone. The intrinsic camera parameters are provided. The camera extrinsics for each frame are estimated by 2D-3D registration against the known 3D model. In an operational setting, we could use a structure-from-motion approach such

Table 4.1: **Reconstruction error [mm] on the C3VD dataset. Surveyed:** points seen at least once. **Extended:** points within 20 mm of a visible point. Anatomical regions: Cecum, Descending, Sigmoid and Transverse. For NeuS, we provide two sets of numbers because the optimization failed on the other sections. In *italics* we mark the sequences where the camera moves less than 1 cm yielding higher errors.

		NeuS		LightNeuS (ours)										
Sequence		**C1a**	**C4b**	**C1a**	**C1b**	**C2a**	**C2b**	**C2c**	**C3a**	**C4a**	**C4b**	**D4a**	**S1a**	**S2a**
Sur.	MedAE	4.53	10.6	0.95	4.85	1.40	3.26	2.57	1.12	1.90	1.41	2.66	4.23	1.19
	MAE	5.07	10.6	1.48	5.11	1.54	3.65	3.00	2.54	2.14	1.63	3.26	4.33	1.89
	RMSE	6.40	11.6	2.01	5.63	1.87	4.39	3.74	5.49	2.92	2.10	4.08	4.96	2.78
Ext.	MedAE	4.68	5.35	0.83	4.89	1.41	3.32	2.54	1.27	1.91	1.45	4.50	4.01	1.40
	MAE	6.24	6.74	1.26	5.10	1.56	3.70	3.01	3.83	2.18	1.72	6.61	4.19	2.36
	RMSE	8.77	8.56	1.72	5.60	1.90	4.42	3.77	7.96	2.95	2.20	9.32	4.87	3.96

LightNeuS (ours)												
S3a	**S3b**	**T1a**	**T1b**	**T2a**	**T2b**	**T4a**	**Mean**	***T2c***	***T3a***	***T3b***	***T4b***	***Mean***
2.57	3.63	3.43	2.33	2.24	2.16	1.15	**2.39**	*5.07*	*6.39*	*11.0*	*1.75*	***6.04***
2.68	4.16	3.47	2.72	2.28	2.30	2.31	**2.80**	*5.45*	*8.65*	*12.1*	*6.70*	***8.23***
3.18	4.81	4.07	3.34	2.58	2.70	3.79	**3.58**	*6.48*	*10.7*	*14.4*	*11.3*	***10.7***
2.87	3.54	3.38	2.69	2.19	2.12	1.29	**2.53**	*4.44*	*6.54*	*13.6*	*8.00*	***8.16***
3.27	4.64	3.31	3.21	2.22	2.28	2.22	**3.15**	*5.36*	*8.10*	*14.1*	*10.4*	***9.47***
4.04	6.10	3.86	3.96	2.55	2.69	3.32	**4.18**	*6.78*	*9.94*	*15.9*	*13.9*	***11.6***

as COLMAP (Schönberger and Frahm, 2016) or a SLAM technique such as Gómez-Rodríguez et al. (2021), which have been shown to work well in endoscopic settings. The gain values were easily estimated from the dataset itself. For vignetting, we use the calibration obtained from a colonoscope of the same brand and series from the EndoMapper dataset (Azagra et al., 2023).

During training, we follow the NeuS paper approach of using a few informative frames per scene, as separated as possible, by sampling each video uniformly. For each sequence, we train both the vanilla NeuS and our LighNeuS using 20 frames each time. They are extracted uniformly over the duration of the video. We use the same batch size and number of iterations as in the original NeuS paper, 512 and 300k respectively. Once the network is trained, we can extract triangulated meshes from the reconstruction. Since the C3VD dataset comprises a ground-truth triangle mesh, we compute point-to-triangle distances from all the vertices in the reconstruction to the closest ground-truth triangle.

In the first rows of Table 4.1, we report median (MedAE), mean (MAE), and root mean square (RMSE) values of these distances for all vertices seen in at least one image. Columns show the result for 22 sequences. We note 18 sequences where the camera moved at least 1 cm, and the reconstruction yielded a mean error of 2.80 mm. The other four smaller trajectories (<1cm) lack parallax and the error is higher (8.23mm).

This is in the range of reported accuracy in the literature for monocular dense non-watertight depth estimation, 1.1 mm in Mahmoud et al. (2018) for high parallax

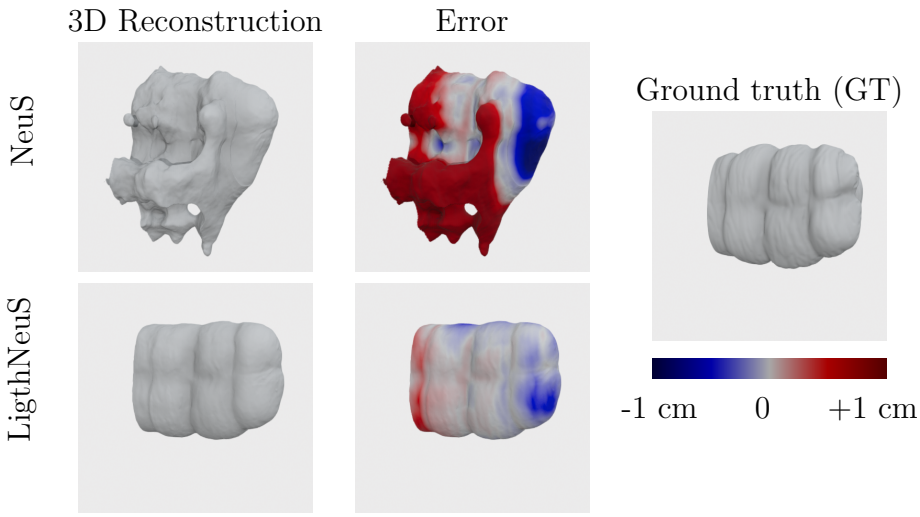


Figure 4.2: **Benefits of illumination decline**. Result on the *"Cecum 1 a"* sequence. **Top:** The NeuS reconstruction exhibits multiple artifacts that make it unusable. **Bottom:** Our reconstruction is much closer to the ground truth shape. The error is shown in blue if the reconstruction is inside the surface, and in red otherwise. A fully saturated red or blue denotes an error of more than 1cm and grey denotes no error at all.

geometry in laparoscopy, which is a much more favorable geometry than the one we have here, or 0.85 mm for the significantly smaller-size cavities of endoscopic endonasal surgery (ESS) (Liu et al., 2022).

In contrast, vanilla NeuS assumes constant illumination. The strong light changes typical of endoscopy fatally mislead the method. We only report numerical results of NeuS in two sequences because in all the rest, the SDF diverges and ends up blown out of the rendering volume, giving no result at all.

We provide a qualitative result in Figure 4.2 and additional ones in section 4.6. Note that the watertight prior inherent to an SDF allows the network to hallucinate unseen areas. Remarkably, these unsurveyed areas continue the tubular shape of the colon and we found them to be mostly accurate when compared to the ground truth. For example, the curved areas of the colon where a wall is occluded behind the corner of the curve is reconstructed, as shown in Figure 4.3. This ability to "fill in" observation gaps may be useful in providing the endoscopist with an estimate of the percentage of unsurveyed area during a procedure.

We hypothesize that this desirable behavior stems from the fact that the network learns an empirical shape prior from the observed anatomy of the colon. However, we don't expect this behavior to hold for distant unseen parts, but only for regions closer than 20 mm to one observation. In the last rows of Table 4.1, we compute accuracy metrics for this *extended* region. It includes not only surveyed areas, but also neighboring areas that were not observed.

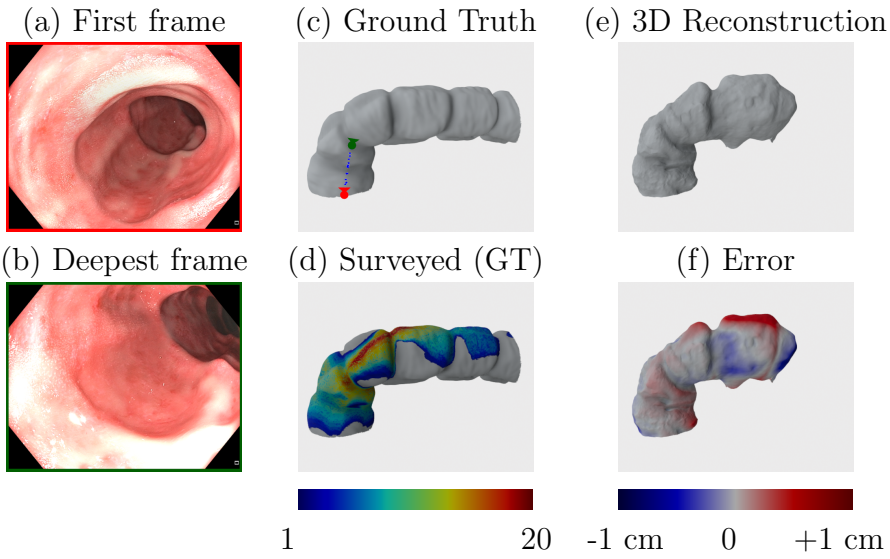


Figure 4.3: **Reconstructing partially observed regions**. Results on *"Transcending 4 a"* sequence. The camera performs a short trajectory from (a) to (b). In (c) we represent both frames and intermediate camera poses. (d) Number of frames seeing each surface point, with GT unobserved areas shown in gray. (e) We managed to reconstruct a curved section of the colon. (f) Our method plausibly estimates the wall of the colon at the right of camera (b), although it was never seen in the images.

4.5 Conclusion

We have presented a method for 3D dense multi-view reconstruction from endoscopic images. We are the first to show that neural radiance fields can be used to obtain accurate dense reconstructions of colon sections of significant length. At the heart of our approach, is exploiting the correlation between depth and brightness. We have observed that, without it, neural reconstruction fails.

The current method could be used offline for post-exploration coverage analysis and endoscopist training. But real-time performance could be achieved in the future as the new NeuS2 (Wang et al., 2023) converges in minutes, enabling automatic coverage reporting. Similar to other reconstruction methods, for now our approach works in areas of the colon where there is little deformation. Several sub-maps of non-deformed areas can be created if necessary. However, this limitation could be overcome by adopting the deformable NeRFs formalism (Park et al., 2021).

4.6 Additional results

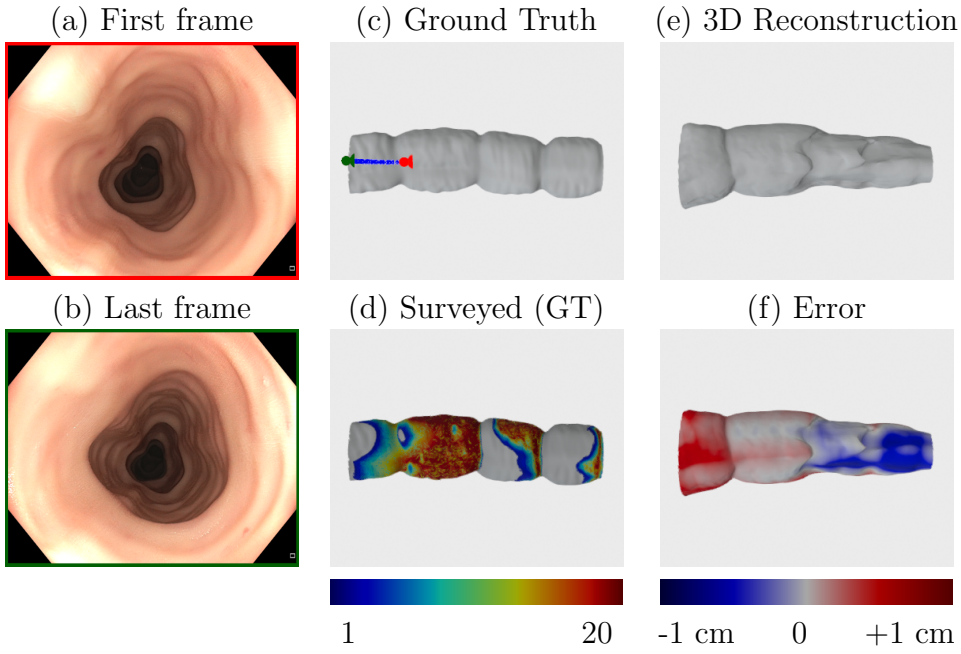


Figure 4.4: **Reconstructing with low parallax in LightNeuS**. Results on *"Transcending 1 a"* sequence. (c) The camera travels in a straight line, covering less than a third of the section. As shown in (a) and (b), the haustra completely hide the background walls. (e) Consequently, the reconstruction underestimates the diameter of the end of the tube. However, the three characteristic folds in our reconstructed colon match the ground truth in number and location. In addition, areas observed multiple times —red in (d)— are reconstructed with high accuracy —gray in (f).

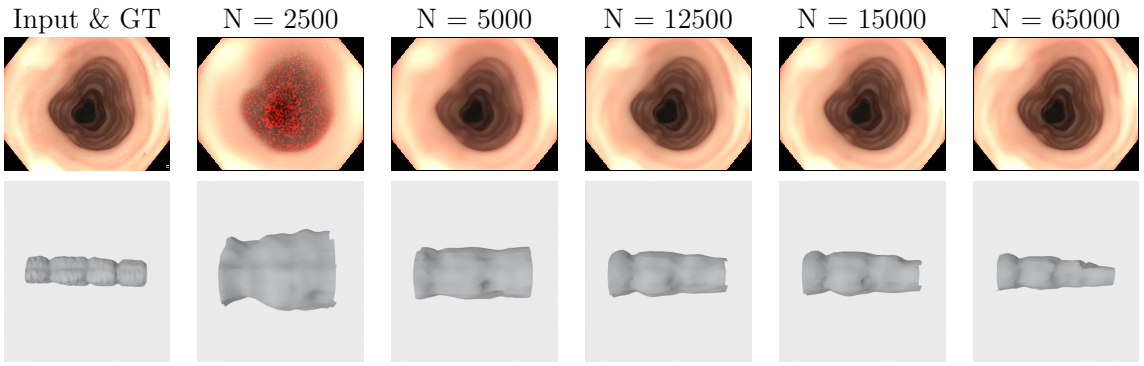


Figure 4.5: **Reconstruction convergence in LightNeuS**. Results on *"Transcending 1 a"* sequence. We show the intermediate results for N optimisation iterations. We see how the reconstruction converges quickly. In 65k iterations we already have a reasonable solution, compared to the 300k iterations proposed by the authors of NeuS.

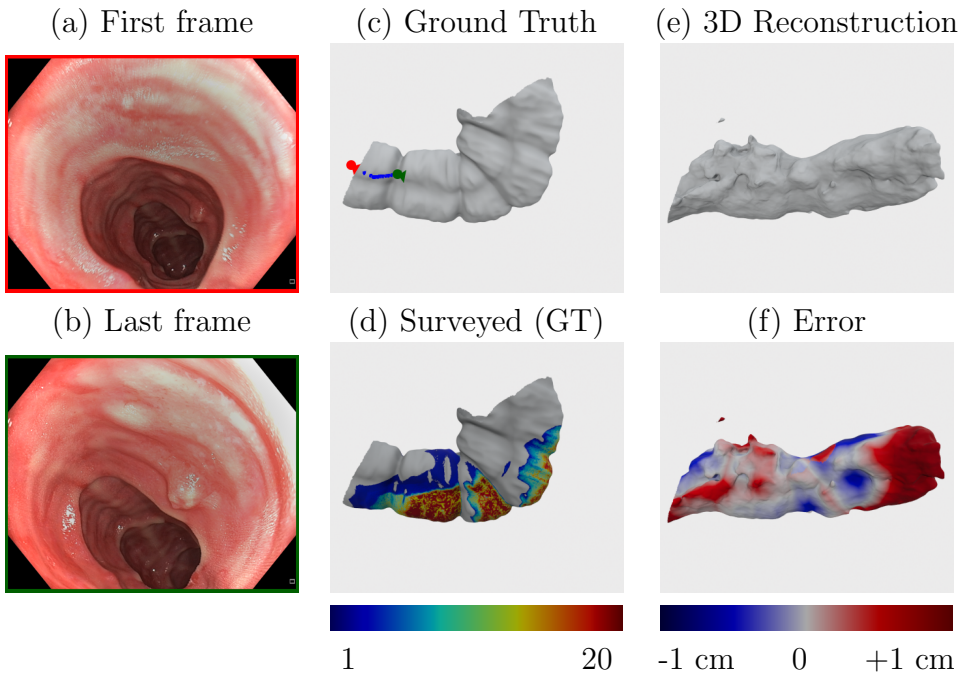


Figure 4.6: **Reconstructing congruent shapes in LightNeuS**. Results on *"Descending 4 a"* sequence. (c) The camera takes the most challenging route: the shortest translation of the sequences tested, turning towards the right wall. (d) This results in very poor coverage, especially to the left of the camera. The curve to the left is never seen. (e) In this way we check that our method only "hallucinates" partially observed areas, based on the structure of the region it has actually observed. As a result, the reconstruction continues as a straight tube. Again, areas observed multiple times —red in (d)— are reconstructed with high accuracy —gray in (f).

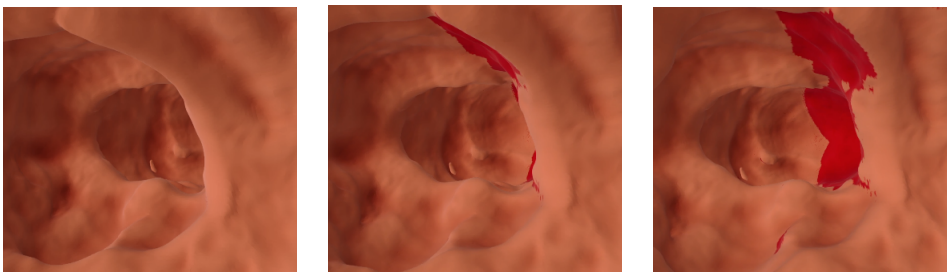


Figure 4.7: **Post-intervention 3D visualization in LightNeuS**. Results on *"Transcending 4 a"* sequence. Our reconstructions would allow physicians to revisit the area explored during the endoscopy. This opens the door to augmented reality (AR) in post-intervention diagnostics. As an example, we show a visualisation of the area not surveyed during the procedure, marked in red. After inspecting a colon section, our watertight surface displays (a) visualized and (b), (c) non-visualized mucosa. The doctor can analyse non-visualized areas and make decisions about subsequent exploration. A video of this demonstration is included in https://youtu.be/YnyUutpGGg4.

Chapter 5

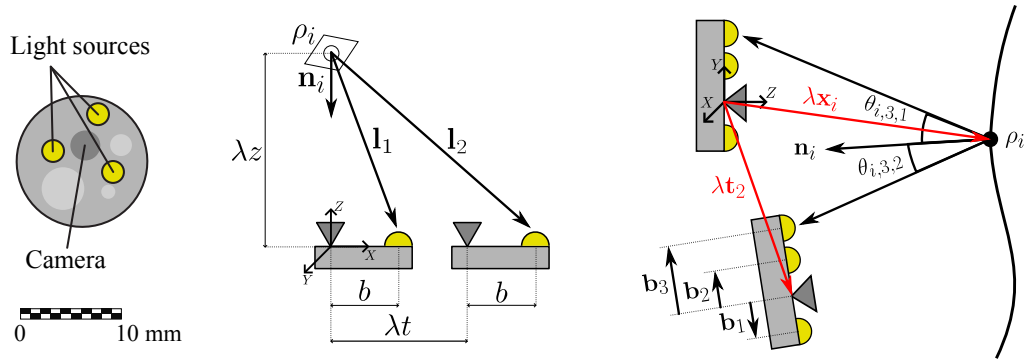
Near-Light Monocular Metric Scale Estimation in Endoscopy

5.1 Introduction

In current endoscopic procedures, endoscope navigation, localization, and tissue measurement are performed manually. Recent advances in Visual Simultaneous Localization and Mapping (VSLAM) for endoscopy (Lamarca et al., 2020; Ma et al., 2021; Gómez-Rodríguez et al., 2024) offer the promise of live 3D reconstructions, that will enable autonomous or assisted navigation and robotized interventions. Most specialties use just monocular endoscopes to reduce bulk and cost. However, using a moving monocular camera, the absolute scale of the environment is unobservable, and the 3D reconstructions and trajectories obtained have an unknown scale factor. This also introduces scale drift which significantly reduces map accuracy.

However, endoscopes are equipped with light sources attached to the camera, which introduce significant illumination variations in the scene. Our key insight is to leverage these illumination changes through near-light photometry to accurately recover the true metric scale of monocular reconstructions. Photometry is scale-dependent due to two factors: the inverse-square decay of illumination with distance from the light source, and the angle between the incident light and the surface normal when the light source is at a small, but nonzero, baseline from the camera's optical center (Figure 5.1).

We demonstrate how true scale can be recovered just from the images captured by a standard monocular endoscope in two steps: first, obtaining an up-to-scale reconstruction with SfM or VSLAM, and then performing photometric optimization to recover scale, gains, and per-point albedo. Our contributions are: (1) A scale-dependent near-light photometric model applicable to any monocular, up-to-scale multi-view reconstruction. (2) A photometric optimization method to estimate true metric scale, per-point albedo, and camera gain. (3) An initialization technique to enhance conver-



(a) Endoscope tip (b) Simplified model (c) Full model

Figure 5.1: Our method estimates the metric scale factor λ by leveraging a near-light illumination model applied to multi-view images captured with a monocular endoscope.

gence and avoid local minima. (4) Simulations and real experiments demonstrating the scale accuracy achievable in endoscopy.

5.2 Related work

Feature-based monocular SLAM (Campos et al., 2021; Elvira et al., 2024) and SfM (Schönberger and Frahm, 2016) recover up-to-scale geometry via bundle adjustment, ignoring photometry. Photometric monocular SLAM (Engel et al., 2017; Zubizarreta et al., 2020) estimates geometry, albedo, and camera gains but assumes constant illumination and cannot recover scale. In contrast, we first estimate up-to-scale geometry via bundle adjustment and then determine metric scale, albedo, and camera gain using non-linear optimization of near-light photometric errors.

Our method relates to photometric stereo, originating from Woodham (1980), which used orthographic images from a fixed camera with distant, switchable light sources to recover surface normals, but not scale. The first method to recover metric scale was Iwahori et al. (1990), leveraging inverse-square illumination decay from multiple point lights at known positions. Near-field photometric stereo, where the camera and light sources are close to the object, was formalized in Mecca et al. (2014), with semi-calibrated approaches introduced by Quéau et al. (2017), requiring known light positions but unknown intensities. Using an endoscope for calibrated near-light photometric stereo was first explored in Collins and Bartoli (2012a), combining three colored light sources into a single shot for true-scale reconstruction. In summary, near-light photometric stereo can generate real-scale reconstructions from three or more images with different point light sources. However, for endoscopy, the device must be modified to control the lights or use colored lights, which is impractical in clinical settings.

Other hardware modifications use structured light to overlay a metric scale on the

image for polyp measurement (Yoshioka et al., 2021; von Renteln et al., 2023). Some works train a deep network to classify polyps into two size classes (smaller or larger than 10 mm) (Itoh et al., 2018, 2021) or to predict dense depth with true scale for polyp measurement (Du et al., 2024; Wang et al., 2024a). However, these methods require task-specific learning and do not generalize to other medical applications. In contrast, we achieve metric-scale reconstruction and estimate the camera's metric trajectory using a standard endoscope without hardware modifications or task-specific learning, relying solely on near-light photometry.

Near-light shape-from-shading was used in Wu et al. (2010) to recover a metric-scale reconstruction from a single perspective image, but relying on strong assumptions of known light intensity and a constant, known albedo. Similar methods (Gonçalves et al., 2015; Batlle et al., 2022) assumed that the light source is located at the optical center, i.e., the baseline is zero. As a result, it becomes impossible to disambiguate scale from albedo. This yielded only up-to-scale reconstructions due to three unknown factors: illumination power, camera gain, and surface albedo. Multi-view near-light photometry to jointly recover scale, gain, and per-point albedo was first proposed in Fernandes-Araujo and Montiel (2022), but only validated through simplistic simulations with up-to-scale ground truth geometry. We take a step further by demonstrating that the method can work from the near-light images alone in realistic simulations and in real colonoscopies.

5.3 Fundamentals

To analyze some properties of near-light photometry we first consider a simplified two-view problem (Fernandes-Araujo and Montiel, 2022, Figure 5.1b). It assumes a moving monocular camera with a single point light source at a distance b from the optical center, observing a Lambertian point with albedo ρ, which lies along the camera's optical axis at a depth λz, with its normal pointing towards the camera. The second camera is translated λt to the right. We aim to compute the unknown scale λ.

Assuming the light intensity L_0 and camera gain g are constant, and no gamma compression, the intensity of the point i in each image k will be (Iwahori et al., 1990):

$$I_{i,k}(\lambda) = \frac{\rho_i g L_0}{\pi} \frac{\mathbf{n}_i \cdot \mathbf{l}_k}{\|\mathbf{l}_k\|^3} = \frac{\rho'_i}{\pi} \frac{\mathbf{n}_i \cdot \mathbf{l}_k}{\|\mathbf{l}_k\|^3} \tag{5.1}$$

where $\rho'_i = \rho_i g L_0$ is a scaled albedo. The intensities in both images are:

$$I_{i,1} = \frac{\rho'_i}{\pi} \frac{\lambda z}{\left(b^2 + \lambda^2 z^2\right)^{3/2}} \quad , \quad I_{i,2} = \frac{\rho'_i}{\pi} \frac{\lambda z}{\left((\lambda t + b)^2 + \lambda^2 z^2\right)^{3/2}} \tag{5.2}$$

We can eliminate ρ'_i to get a second-order equation on λ:

$$b^2 + \lambda^2 z^2 = c\left((\lambda t + b)^2 + \lambda^2 z^2\right) \tag{5.3}$$

where $c = (I_{i,2}/I_{i,1})^{2/3}$ is a known constant, determined form the intensities measured in the images. This equation allows to obtain λ, except when $b = 0$, in which case λ simplifies away, and cannot be solved.

We can conclude that in a multi-view near-light scenario, metric scale is observable only when the baseline between camera and light source is non-zero, even if light intensity, gain and albedo are unknown. Also, we can expect scale accuracy to degrade when the distance to the surface is too large compared with the camera-light baseline. The remainder of the chapter proposes a practical method for obtaining the scale in real endoscopic settings and studies its accuracy.

5.4 EndoMetric

Our proposed method, EndoMetric, works in two steps: first obtain an *up-to-scale multi-view reconstruction* of the scene using any SLAM or SfM method, and then solve an optimization problem for *metric scale estimation* that impose albedo consistency. The key for scale estimation is leveraging a *near-light photometric model* that takes into account the baseline of the light sources with respect to the camera and the inverse-square law of illumination decline with distance.

5.4.1 Up-to-scale Multi-view Reconstruction

Classical multi-view geometry can produce, from a sequence of calibrated monocular images (at least two), the geometry of n scene points $\{\mathbf{x}_i\}_{i=1}^n$ and m camera poses $\{\mathbf{T}_k = (\mathbf{R}_k, \mathbf{t}_k)\}_{k=1}^m$ up to an unknown scale factor λ, by solving bundle adjustment, i.e. the non-linear optimization of the re-projection errors of the points matched along the sequence. We use the well known COLMAP (Schönberger and Frahm, 2016) to compute multiview geometry from endoscopic sequences. Photometric models need not only the sparse scene geometry, but also the surface normals at each scene point. We propose to compute the normals $\mathbf{n}_i$ by fitting a plane to the p neighbors of each scene point.

5.4.2 Near-light Photometric Model

We assume a calibrated mobile camera with r fixed point light sources at known positions relative to the optical center $\{\mathbf{b}_j\}_{j=1}^r$. All lights share the same intensity L_0, uniformly distributed in all directions, with a calibrated lens vignetting. Points near

specular reflections are discarded, and the surface is assumed to be Lambertian with an unknown, varying albedo.

We have m grayscale images $\{I_k\}_{k=1}^{m}$ taken from m poses while observing n scene points. For a given point i in an image k the image intensity depends on the point albedo ρ_i, the camera gain g_k and the light intensity L_0. The observed albedos are always multiplied by the camera gain, g_k, and L_0. As these values are not provided by currently available endoscopes, we define two new variables that are observable from the images:

$$\rho'_i = \rho_i g_1 L_0 \quad , \quad g'_k = \frac{g_k}{g_1} \tag{5.4}$$

where ρ'_i is a scaled albedo (that may be greater than 1) and g'_k represents the gain change with respect to the first image.

Illumination depends on the incidence angle and the inverse-square of the distance between light and point. Then, the perceived radiance depends on the unknown scale factor λ of the multi-view reconstruction as (Figure 5.1c):

$$\mathcal{I}_{i,k}(\rho'_i, g'_k, \lambda) = \left(\frac{\rho'_i g'_k}{\pi} \sum_{j=1}^{r} \frac{\cos\theta_{i,j,k}(\lambda)}{\|\lambda \mathbf{x}_i - (\mathbf{R}_k \ \mathbf{b}_j + \lambda \mathbf{t}_k)\|^2} V(\mathbf{x}_i) \right)^{1/\gamma} \tag{5.5}$$

where $\cos\theta_{i,j,k}(\lambda) = \mathbf{n}_i \frac{\lambda \mathbf{x}_i - (\mathbf{R}_k \ \mathbf{b}_j + \lambda \mathbf{t}_k)}{\|\lambda \mathbf{x}_i - (\mathbf{R}_k \ \mathbf{b}_j + \lambda \mathbf{t}_k)\|}$, $V(\mathbf{x}_i)$ is the calibrated lens vignetting and γ is gamma compression.

Note that, if all light-camera baselines $\mathbf{b}_j$ were zero, λ^2 could be extracted from the denominator and all intensities would be proportional to ρ'_i/λ^2, producing a fundamental ambiguity: you can multiply the scale by any constant c just multiplying all albedos by c^2. So, also in the general case, a non-zero baseline is needed to break the ambiguity and estimate true scale and per-point albedo.

5.4.3 Metric Scale Estimation

Given the up-to-scale reconstruction, our near-light photometric model, and the original images $I_1 ... I_m$, our goal is to recover the scale factor λ and, as a side product, the point albedos ρ'_i and camera gain changes g'_k. This can be achieved by minimizing the photometric error with respect to the model:

$$\underset{\{\lambda, \rho'_1 ... \rho'_n, g'_2 .. g'_m\}}{\arg\min} \sum_{i,k} \|\mathcal{I}_{i,k}(\rho'_i, g'_k, \lambda) - I_{i,k}\|_\epsilon^2 \tag{5.6}$$

where $I_{i,k}$ is the intensity of point i observed in image I_k. A robust cost function is used to reduce the influence of spurious data. This nonlinear optimization is solved using the Levenberg-Marquardt method implemented in Ceres (Agarwal et al., 2023).

5.4.4 Initial Guess for the Scale

To avoid local minima and achieve faster convergence it is crucial to find good initial values for the optimization variables (λ, ρ'_i, and g'_k). Indeed, these variables are closely related, therefore, instead of estimating their initial values separately as in Fernandes-Araujo and Montiel (2022), we propose to estimate them jointly.

We perform an exhaustive search for the scale parameter λ over a logarithmic space Λ. For each trial value $\hat{\lambda}$, we estimate the albedo values $\hat{\rho}'_i$ solving (5.5) for each point in the first image. Then, to estimate the relative gains of the rest of images, we perform a robust regression. First, we undo the gamma compression I^γ to work in linear space. Next, we find the gain value $\hat{g}_k$ that minimizes the difference between the real value $I^\gamma_{i,k}$ and the estimated value $\mathcal{I}^\gamma_{i,k}$ of the points using the robust cost function. Finally, we select the value of $\hat{\lambda}$ with the lowest residual according to (5.6).

5.5 Experiments

5.5.1 Datasets

Simulation dataset. Real endoscopic images with a ground-truth metric scale are difficult to obtain. Therefore, we assessed the accuracy of our method through simulations. Our endoscope has a monocular fisheye camera and three surrounding light sources with a ∼3mm baseline (Figure 5.1a). We used a real 3D mesh from İncetan et al. (2021), a Lambertian illumination model, and Gaussian pixel noise of 4 gray levels. Examples are available in https://youtu.be/bjRvyTS7_CI.

EndoMapper dataset (Azagra et al., 2023). We validate our method using real endoscopy videos to assess its performance in real-world conditions. To our knowledge, EndoMapper is the only dataset that provides both photometric calibration of the endoscope and metric-scale annotations of polyp sizes estimated by endoscopists. It uses a calibrated Olympus endoscope with three light sources. We interpret the provided light spread as vignetting, since both were jointly estimated, and we use isotropic light sources positioned according to the manufacturer's datasheet.

5.5.2 Impact of Distance to the Surface

In our *simulation dataset*, the endoscope is positioned to face a polyp at varying distances from the surface (3 to 20 mm), capturing images from slightly different viewpoints, which could be easily achieved in practice by bending the endoscope tip. Four images are used at a time to reconstruct the 3D shape with COLMAP and estimate

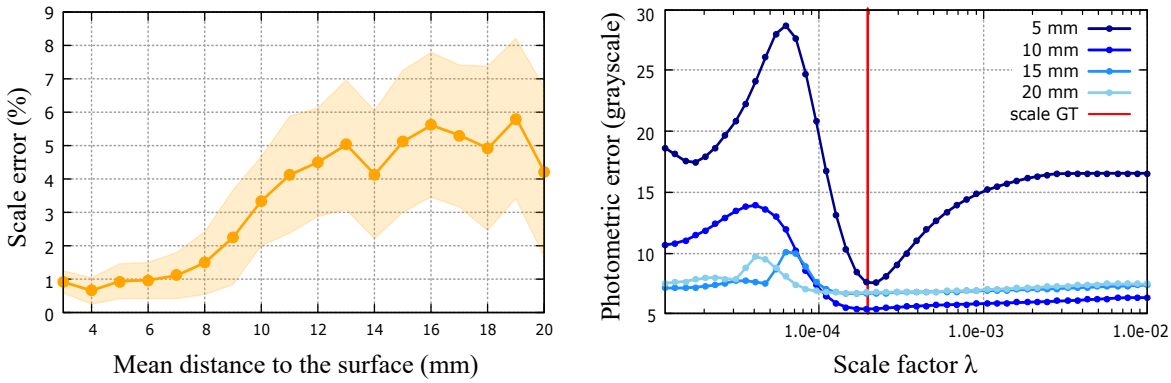


Figure 5.2: Method accuracy depends on surface distance, performing best when it matches the light-camera baseline. *Left:* Scale error increases from 1% to 5% with distance. *Right:* Greater distances weaken the photometric cost function's minimum.

Table 5.1: Accuracy at near-field ranges (5mm). The study highlights the importance of precise multi-view geometry $\mathbf{T}_k$ and the benefits of our initial guess.

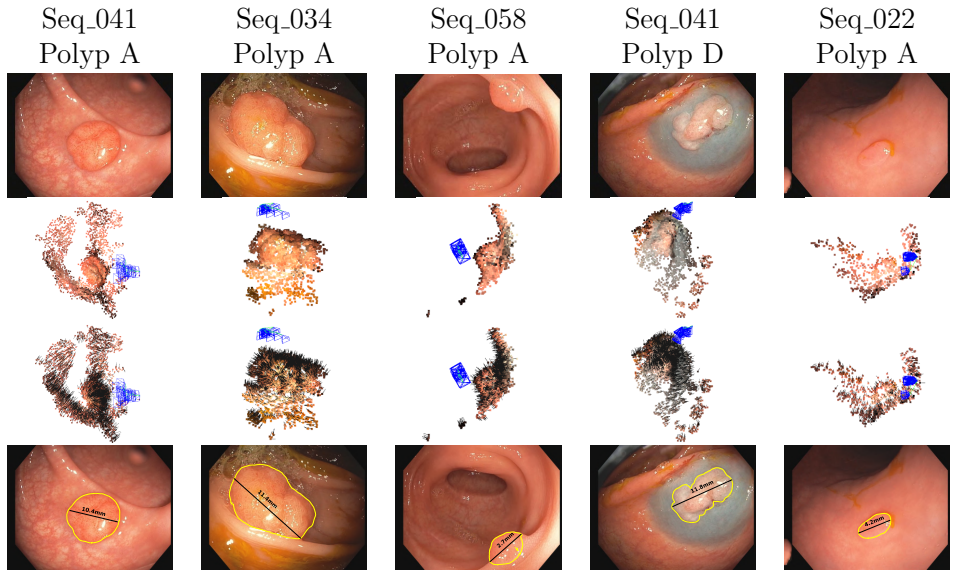
		Optimized			Initial guess	$\mathbf{T}_k$	$\mathbf{n}_i$	% error		
		λ	ρ'	g'				λ	ρ'	g'
EndoMetric		✓	✓	✓	✓	SfM	SfM	0.95	5.48	3.37
Ablation study	A	✓	✓	✓	✓	SfM	**GT**	0.81	4.52	4.04
	B	✓	✓	**GT**	✓	SfM	GT	0.80	3.99	–
	C	✓	✓	GT	✓	**GT**	GT	0.17	3.44	–
	D	✓	✓	✓	✗	SfM	SfM	38.21	18.43	3.88
	E	✓	✓	✓	(Fernandes, 2022)	GT	GT	1.23	4.25	4.57

its scale. Figure 5.2L presents the average scale error as a function of distance to the surface. Since different image sets may yield varying results, we report the mean and standard deviation over five experiments.

The method achieves an error of approximately 1% at distances up to 8 mm, which are typical in endoscopic procedures. This confirms its effectiveness in near-field conditions, where the distances between the camera, light sources, and surface are of the same order of magnitude. The scale error rises to 5% as we move away from the surface, since the photometric cost function exhibits a less marked minimum (Figure 5.2R).

5.5.3 Impact of Multi-View Reconstruction Accuracy

Focusing on near-field conditions, Table 5.1 presents an ablation study using only images from our *simulation dataset* captured at approximately 5 mm from the surface. Results show that scale accuracy remains similar whether ground-truth surface normals (row A) or camera gain values (row B) are provided. However, the error decreases significantly in the scenario with the most ground-truth information (row C), suggesting that improving the underlying multi-view reconstruction could enhance the accuracy of our scale estimation method.



	Seq_041 Polyp A	Seq_034 Polyp A	Seq_058 Polyp A	Seq_041 Polyp D	Seq_022 Polyp A
Our method	10.4mm	11.4mm	2.7mm	11.8mm	4.2mm
Endoscopist	10mm	10mm	3mm	8–10mm	4–5mm
Discrepancy	0.4mm (4%)	1.4mm (14%)	0.3mm (10%)	2.8mm (31%)	0.3mm (7%)

Figure 5.3: Results in EndoMapper dataset (Azagra et al., 2023).

5.5.4 Impact of Initial Guess

Our photometric cost function is non-linear and can exhibit multiple local minima, as illustrated in Figure 5.2R. To ensure robustness and generality, we propose a method for computing the initial seed for non-linear optimization. Without this step, our method may get trapped in a sub-optimal solution, resulting in large errors, as shown in Table 5.1 (row D). Previous work Fernandes-Araujo and Montiel (2022) initialized albedo and gains to constant values and assumed known camera poses and scene geometry. We tested this configuration with our software and achieved an error of 1.23% in estimating the real scale under these ideal conditions (row E). In contrast, despite estimating camera poses and scene geometry automatically, our method obtains a smaller error of 0.95%, thanks to a better initial guess.

5.5.5 Real Polyps Measurement

In the *EndoMapper dataset*, some sequences include metadata with annotations derived from the endoscopist's speech recorded during the procedure. We selected five polyps where the practitioner estimated the lesion size. The selection criteria included a clear view of the polyp, a distance to the surface of 5–15mm, and smooth endoscope tip motion. We use four seconds of video at 5 FPS—yielding a total of 20 non-contiguous frames—and process them with COLMAP to reconstruct the 3D shape, whose scale is computed by our method. We apply Segment Anything (Kirillov et al., 2023) to

identify the polyp in a single frame. The size of the poly is determined by measuring the longest diameter of the 3D points within the segmented region.

On an i7 10700k 3.8 GHz CPU, the running time is 45 s for COLMAP reconstruction, 15 s for the initial scale estimation (Python prototype) and 0.4 s for minimizing eq. (5.6) with Ceres. In future work, we will replace COLMAP with a real-time SLAM method, and optimize the initial scale estimation in C++.

Figure 5.3 presents an input image, the reconstructed 3D shape, the estimated surface normals, and our diameter measurement within the polyp boundaries. On average, our measurements deviate from the endoscopist's estimation by 1.0 mm (13%), demonstrating the potential of the method for standardizing polyp size assessment.

5.6 Conclusions

We have presented, for the first time, a method to obtain 3D reconstructions with real metric scale from a conventional monocular endoscope, under unknown varying albedo, solely based on physical principles. The method does not require any application-specific learning, prior knowledge, or hardware modifications—only a calibrated light-camera setup. Our simulations demonstrate that accurate metric scale can be recovered under practical conditions. Our experiments on the EndoMapper dataset show that the method produces polyp measurements closely matching those estimated visually by an endoscopist. This provides a preliminary, yet solid, proof that the method bridges the simulation-to-real gap and effectively estimates metric scale in real images. A quantitative evaluation of its accuracy will require a dataset with ground-truth polyp size annotations.

Near-field photometry paves the way for real-scale visual SLAM with monocular endoscopes. This will be critical in the short term for accurate measurements, and in the long term, for autonomous robotic exploration and surgery.

Chapter 6

Conclusions and Future Work

This thesis lays the foundations for exploiting lighting cues in standard clinical endoscopy. As a result, it enables *dense, metric, monocular* SLAM and 3D reconstruction without altering the widely adopted medical devices. Central to this effort is a photometric model for endoscopic imaging, that accounts for *illumination decline*. By embracing rather than ignoring the brightness variations caused by near-field lighting, this work shifts the paradigm—transforming a long-standing challenge into a valuable source of geometric information.

The calibrated photometric model serves as the cornerstone of our contributions, supporting a family of 3D reconstruction techniques that recover depth, reconstruct implicit surfaces, and ultimately achieve metric-scale accuracy with only millimeter-level error. Notably, we introduce a novel self-supervised paradigm that eliminates the need for ground-truth data, which is scarce in medical imaging, and avoids the limitations of learned priors and domain adaptation.

In a context of rapidly emerging reconstruction methods, we extend neural radiance field (NeRF) techniques to colonoscopy by integrating photometric modeling into implicit surface reconstructions. Our method yields watertight models from monocular videos, offering a compelling tool for preoperative planning and automatic assessment of mucosal coverage in colorectal cancer screening.

The thesis culminates with a method that breaks through the core limitation of monocular SLAM: the inability to infer metric scale. By relying solely on our physically grounded model of light attenuation and camera-light baselines, we recover scale without the need for learned priors or additional sensors, opening the door to accurate measurement and navigation capabilities using standard endoscopic video.

This pioneering step in leveraging illumination decline for monocular SLAM in endoscopy demonstrated that a near-field light source can be exploited to recover geometric information. It has since inspired follow-up research by other groups, including methods that integrate near-field illumination into traditional Bundle Adjustment (BA)

for semi-dense SLAM (Beltran et al., 2024), and others that further disentangle albedo, depth, and shading to advance non-Lambertian image decomposition (Daher et al., 2025).

Thanks to this thesis, several promising avenues for future work have opened up. EndoMetric, the only sparse technique presented, could be extended to support dense reconstruction. Additionally, most of the methods described here assume locally rigid or static anatomy, but could be adapted to handle deformable, dense models—greatly expanding their applicability in clinical settings. Another important direction is the integration of photometrically informed depth cues into real-time, metric SLAM frameworks, which remains an open and impactful challenge critical to the long-term goal of autonomous endoscopy.

The prospect of providing endoscopists with real-time, 3D anatomical visualization and precise metric measurements via augmented reality is becoming more feasible. We posit that exploiting illumination decline will be vital to bring this future to life.

Conclusiones y trabajo futuro

Esta tesis sienta las bases para aprovechar la iluminación en la endoscopia clínica estándar. Como resultado, permite la reconstrucción 3D y el SLAM *monocular, denso y métrico* sin necesidad de alterar los dispositivos médicos existentes. Un elemento central de este trabajo es el modelo fotométrico para imágenes de endoscopia, que tiene en cuenta la *decaimiento de la luz*. Al aprovechar las variaciones de brillo causadas por la iluminación en lugar de ignorarlas, esta tesis cambia el paradigma y transforma un viejo desafío en una valiosa fuente de información geométrica.

El modelo fotométrico calibrado es la piedra angular de nuestras contribuciones, ya que respalda una familia de técnicas de reconstrucción 3D que recuperan la profundidad, reconstruyen superficies implícitas y, en última instancia, logran recuperar la escala métrica con un error de solo milímetros. En particular, introducimos un nuevo paradigma autosupervisado que elimina la necesidad de datos etiquetados, que son escasos en las imágenes médicas, y evita limitaciones como la dependencia de información aprendida a priori o la difícil adaptación a cambios de dominio.

En un mundo de métodos de reconstrucción que evolucionan rápidamente, esta tesis amplía las técnicas de *neural radiance fields* (NeRF) a la colonoscopia, integrando nuestro modelo fotométrico con el aprendizaje de superficies implícitas. Nuestro método produce modelos estancos a partir de vídeos monoculares, que son útiles, por ejemplo, en la planificación preoperatoria o en la evaluación automática del porcentaje de mucosa observada durante el cribado del cáncer de colon.

La tesis culmina con un método que supera la limitación fundamental del SLAM monocular: la incapacidad de percibir la escala métrica. Basándonos únicamente en nuestro modelo físico del decaimiento de la luz y en la distancia entre la cámara y la luz, recuperamos la escala sin necesidad de utilizar información a priori ni sensores adicionales. Esto abre la puerta a la navegación y la medición precisa con un videoendoscopio estándar.

Este paso pionero en el uso del decaimiento de la iluminación para SLAM monocular en endoscopia muestra que se puede aprovechar una fuente de luz para recuperar información geométrica. Esta idea ha inspirado investigaciones por parte de otros grupos,

como un método que integra la iluminación en la optimización tradicional (*Bundle Adjustment* o BA) para SLAM semidenso (Beltran et al., 2024) y otro que avanza aún más en la desambiguación del albedo, la profundidad y el transporte de luz para descomponer imágenes no lambertianas (Daher et al., 2025).

Gracias a esta tesis, se han abierto varias vías prometedoras de trabajo futuro. EndoMetric, la única técnica dispersa presentada, podría ampliarse para admitir reconstrucción densa. Además, la mayoría de los métodos aquí descritos asumen una anatomía localmente rígida o estática, pero podrían adaptarse para manejar modelos deformables y densos, lo que ampliaría enormemente su aplicabilidad en entornos clínicos. Otra dirección importante es la integración de profundidad fotométrica en sistemas SLAM métricos en tiempo real, lo que sigue siendo un reto abierto con un gran impacto, fundamental para el objetivo a largo plazo de la endoscopia autónoma.

La perspectiva de ofrecer a los endoscopistas visualización anatómica 3D en tiempo real o mediciones métricas precisas con anotaciones de realidad aumentada es cada vez más real. Postulamos que modelar el decaimiento de la iluminación será fundamental para hacer realidad este futuro.

Appendices

Appendix A

Open-Source Photometric Calibration Software

Source code available at

`https://github.com/endomapper/EM_Dataset-PhotometricCalibration`

This appendix provides a brief overview of the open-source software developed for photometric calibration, which is available on GitHub. The software is designed to facilitate the calibration of endoscopic cameras using the photometric models described in this thesis.

A.1 Overview

The calibration procedure consists of recording a video sequence of a calibration target using the endoscope in a controlled environment, with no external light sources other than those of the endoscope itself. See Figure A.1 for an example. The entry point for the software is the `./run` file, which provides an interactive command-line interface.

Once the video is provided, the software automatically processes it to estimate the photometric parameters of the endoscope by solving a nonlinear least-squares problem with bounds. The software is implemented in Python and uses the `scipy` library for optimization. See `calibration/test_hculb.py` for the main calibration logic.

Two camera models are supported: the `kb4` model, which implements fish-eye distortion according to Kannala and Brandt (2006), and the `poly2` model, which is a pinhole camera with barrel and cushion distortion. The intrinsics and extrinsics of the camera are estimated using Vicalib (Heckman et al., 2016). Camera models are implemented in the `calibration/cameras.py` file.

A.2 Calibration options

The software provides several options for calibration, which can be selected through the `calibration/config_globals.py` file. The available options include:

A.2.1 `OPTIMIZE_LIGHT`: Light source model

This option allows the user to choose between several complexity levels for the light source model. Each of the following model parameters can be optimized or fixed during calibration:

- σ_0 is the light source intensity. We denote as *normalized* (N) the case where this parameter is fixed to 1.
- $\mathbf{P}$ or $\mathbf{x}_l$ is the light source position. We denote as *fixed position* (F) the case where this parameter is fixed, usually to the camera's optical center.
- R or μ is the light spread. We denote as *spot light source* (SLS) the case where the light spread has a radial falloff, and as *point light source* (PLS) the case where the light spread is isotropic.
- $\mathbf{D}$ or $\mathbf{d}_l$ is the light source principal direction of a spot light source (SLS). We denote as *z-oriented* (Z) the case where this parameter is fixed to the camera's forward direction.

All this logic is implemented in the `calibration/lights.py` file, which is based on the definitions by Modrzejewski et al. (2020). We use `SINGLE_NZFSLS` in our baseline and LightNeuS methods, which is a *normalized* (N), *z-oriented* (Z), and *fixed position* (F) *spot light source* (SLS) model. This model assumes a single light source with a known position and orientation, only the light spread is optimized. The software's default value `SINGLE_NSLS` is used in LightDepth and the EndoMapper dataset. Future work could use a more complex `TRI_NFSLS` model to improve the calibration of multiple light sources in EndoMetric.

A.2.2 `OPTIMIZE_BRDF`: Surface reflectance model

This option allows the user to choose between several Bidirectional Reflectance Distribution Functions (BRDF) for the calibration target. The available models are:

- `DIFFUSE`: A simple Lambertian reflectance model, which assumes a uniform surface reflectance.

- `PHONG`: A Phong reflectance model, which includes specular highlights and is suitable for shiny surfaces.
- `LUT`: A lookup table model, which allows for more complex surface reflectance properties.

All this logic is implemented in the `calibration/brdfs.py` file. We used the `LUT` model in the calibration for our baseline method. Then we switched to the `DIFFUSE` model for the rest of the methods, by recording the EndoMapper dataset with a nearly-Lambertian calibration target made of unidirectional carbon fiber.

A.2.3 `OPTIMIZE_VIGNETTING`, `_GAIN`: Camera response model

These options allow the user to fix or optimize some properties of the camera response model. The available options are:

- `VIGNETTING`: This option allows the user to optimize the vignetting effect of the lenses. The value `COSINE` assumes a cosine falloff of the brightness towards the corners of the image. `LUT` uses a lookup table to model the vignetting effect and `NONE` assumes no vignetting (this is the default behavior). See implementation in the `calibration/vignettings.py` file.
- `GAIN`: This option refers to the automatic gain control (AGC) of the camera. It can be estimated in absolute values for all images (`ALL`) or relative to the first frame (`EXCLUDE_FIRST`). The default value is `ALL`, especially when using normalized light sources.

We assume a gamma correction of 2.2 for the camera response, which is a common value for endoscopic cameras. This is implemented in `calibration/cameras.py`.

A.3 Calibration output

The calibration software outputs a XML file with the estimated parameters of the endoscope. An endoscope's `<rig>` tag may have one or more `<camera>` tags, associated with one or more `<light>` sources. Currently, only a single camera and a single virtual light are supported. See Listing A.1 and Figure A.2 for an example.

Each camera tag has a particular `<camera_model>`. This model has a single parameter, the value of the gamma γ response function. By default, vignetting is not estimated (`NONE`). Regarding the light source, the `<light_model>` has four parameters: the intensity value σ_0, the light spread factor μ and two vectors for the light centre $\mathbf{P}$ and the principal direction $\mathbf{D}$.

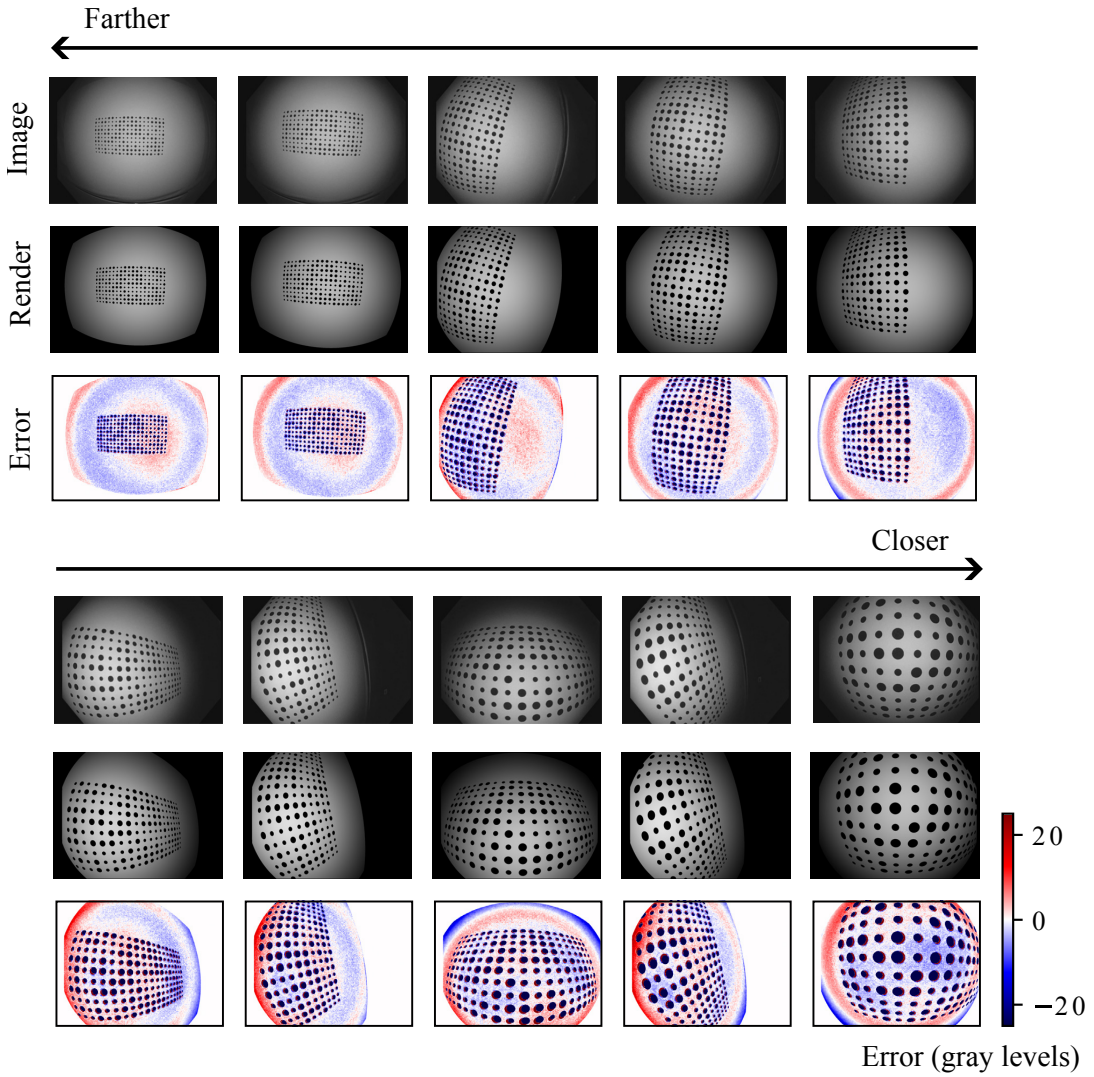
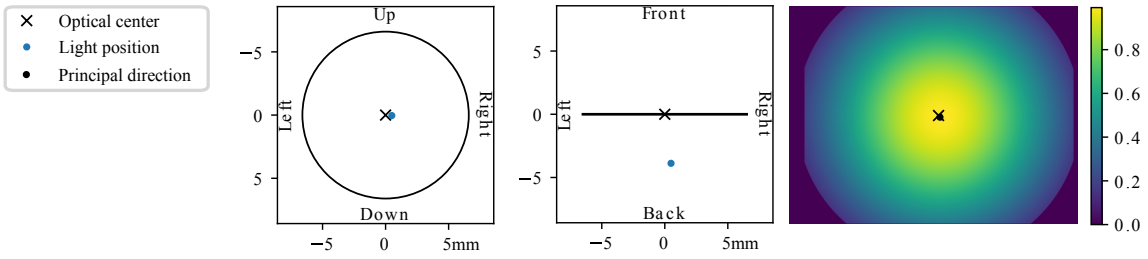


Figure A.1: Example of photometric calibration error on the EndoMapper dataset. The top row shows the input validation frames; the middle and bottom rows display the estimated renderings and the corresponding error maps, respectively.



(a) Endoscope tip front view

(b) Endoscope tip top view

(c) Endoscope vignetting and light spread

Figure A.2: Example of the calibration output for the EndoMapper dataset. The two images on the left show the estimated position of the light source relative to the camera's optical center. The light source is located approximately four millimeters behind the camera.

```
<rig>
  <camera>
    <camera_model name="Endoscope_04" type="gamma" version="1.0">
      <!-- Camera response model -->
      <gamma> [ 2.2 ] </gamma>
    </camera_model>
  </camera>
  <light>
    <light_model type="sls">
      <!-- Spot Light Source (SLS) model
        as in [Modrzejewski et al. (2020)] -->
      <!-- main intensity value -->
      <sigma> 1.000000 </sigma>
      <!-- spread factor -->
      <mu> 3.069096 </mu>
      <!-- light centre in camera reference (3D point) -->
      <P> [ 0.000494; 3.8e-05; -0.00388 ] </P>
      <!-- principal direction in camera reference -->
      <D> [ 0.01028; 0.0115; 0.999881 ] </D>
    </light_model>
  </light>
</rig>
```

Listing A.1: Example of an output XML file generated for the EndoMapper dataset.

Bibliography

S. Agarwal, K. Mierle, and The Ceres Solver Team. Ceres Solver, 2023. URL `https://github.com/ceres-solver/ceres-solver`.

M. A. Armin, G. Chetty, H. De Visser, C. Dumas, F. Grimpen, and O. Salvado. Automated visibility map of the internal colon surface from colonoscopy video. *Int. J. of Computer Assisted Radiology and Surgery*, 11:1599–1610, 2016.

P. Azagra, C. Sostres, Á. Ferrandez, L. Riazuelo, C. Tomasini, O. L. Barbed, J. Morlana, D. Recasens, V. M. Batlle, J. J. Gómez-Rodríguez, R. Elvira, J. López, C. Oriol, J. Civera, J. D. Tardós, A. C. Murillo, A. Lanas, and J. M. M. Montiel. Endomapper dataset of complete calibrated endoscopy procedures. *Scientific Data*, 10(1):671, 2023.

G. Bae, I. Budvytis, C.-K. Yeung, and R. Cipolla. Deep multi-view stereo for dense 3D reconstruction from monocular endoscopic video. In *Int. Conf. on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 774–783. Springer, 2020.

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

V. M. Batlle. *Real-scale 3D reconstruction from monocular endoscope images.* MSc thesis, EINA, Universidad de Zaragoza, 2021.

V. M. Batlle, J. M. M. Montiel, and J. D. Tardós. Photometric single-view dense 3D reconstruction in endoscopy. In *IEEE Int. Conf. on Intelligent Robots and Systems (IROS)*, pages 4904–4910, 2022.

V. M. Batlle, J. M. M. Montiel, P. Fua, and J. D. Tardós. LightNeuS: Neural surface reconstruction in endoscopy using illumination decline. In *Int. Conf. on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 502–512. Springer, 2023a.

V. M. Batlle, J. M. M. Montiel, and J. D. Tardós. EM Dataset: Photometric calibration, 2023b. URL `https://github.com/endomapper/EM_Dataset-PhotometricCalibration`.

A. D. Beltran, D. Rho, M. Niethammer, and R. Sengupta. NFL-BA: Improving endoscopic SLAM with near-field light bundle adjustment. *arXiv preprint arXiv:2412.13176*, 2024.

S. F. Bhat, I. Alhashim, and P. Wonka. Adabins: Depth estimation using adaptive bins. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 4009–4018, 2021.

T. L. Bobrow, M. Golhar, R. Vijayan, V. S. Akshintala, J. R. Garcia, and N. J. Durr. Colonoscopy 3D video dataset with paired depth from 2D-3D registration. *Medical Image Analysis*, 90:102956, 2023. ISSN 1361-8415.

M. Boss, V. Jampani, R. Braun, C. Liu, J. Barron, and H. P. Lensch. Neural-PIL: Neural pre-integrated lighting for reflectance decomposition. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 10691–10704, 2021.

A. Boulch and R. Marlet. Deep learning for robust normal estimation in unstructured point clouds. In *Computer Graphics Forum*, volume 35, pages 281–290, 2016.

C. Campos and J. D. Tardós. Scale-aware direct monocular odometry. In *IEEE Int. Conf. on Intelligent Robots and Systems (IROS)*, 2022.

C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. M. Montiel, and J. D. Tardós. ORB-SLAM3: An accurate open-source library for visual, visual-inertial, and multimap SLAM. *IEEE Transactions on Robotics*, 37(6):1874–1890, 2021.

G. Chen, K. Han, B. Shi, Y. Matsushita, and K.-Y. K. Wong. Self-calibrating deep photometric stereo networks. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 8739–8747, 2019a.

R. J. Chen, T. L. Bobrow, T. Athey, F. Mahmood, and N. J. Durr. SLAM endoscopy enhanced by adversarial depth prediction. *KDD Workshop on Applied Data Science for Healthcare*, 2019b.

Y. Chen, C. Schmid, and C. Sminchisescu. Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera. In *IEEE Int. Conf. on Computer Vision (ICCV)*, pages 7063–7072, 2019c.

K. Cheng, Y. Ma, B. Sun, Y. Li, and X. Chen. Depth estimation for colonoscopy images with self-supervised learning from videos. In *Int. Conf. on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 119—-128. Springer, 2021.

G. Ciuti, M. Visentini-Scarzanella, A. Dore, A. Menciassi, P. Dario, and G.-Z. Yang. Intra-operative monocular 3D reconstruction for image-guided navigation in active locomotion capsule endoscopy. In *IEEE RAS & EMBS Int. Conf. on Biomedical Robotics and Biomechatronics (BioRob)*, pages 768–774, 2012.

J. J. Clark. Active photometric stereo. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 92, pages 29–34, 1992.

T. Collins and A. Bartoli. 3D reconstruction in laparoscopy with close-range photometric stereo. In *Int. Conf. on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 634–642. Springer, 2012a.

T. Collins and A. Bartoli. Towards live monocular 3D laparoscopy using shading and specularity information. In *Int. Conf. on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 11–21. Springer, 2012b.

A. Corona-Figueroa, J. Frawley, S. B. Taylor, S. Bethapudi, H. P. H. Shum, and C. G. Willcocks. MedNeRF: Medical neural radiance fields for reconstructing 3D-aware CT-projections from a single X-ray. In *Int. Conf. of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 3843–3848, 2022.

G. J. Criner, R. Eberhardt, S. Fernandez-Bussy, D. Gompelmann, F. Maldonado, N. Patel, P. L. Shah, D.-J. Slebos, A. Valipour, M. M. Wahidi, M. Weir, and F. J. Herth. Interventional bronchoscopy. *American journal of respiratory and critical care medicine*, 202(1):29–50, 2020.

R. Daher, F. Vasconcelos, and D. Stoyanov. SHADeS: Self-supervised monocular depth estimation through non-lambertian image decomposition. *Int. J. of Computer Assisted Radiology and Surgery*, pages 1–9, 2025.

A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse. MonoSLAM: Real-time single camera SLAM. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1052–1067, 2007.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

S. Du, Q. Zhang, Z. Zhang, C. Cai, X. Li, and D. Qian. Polyp size estimation by generalizing metric depth estimation and monocular 3D reconstruction. In *IEEE Int. Symp. on Biomedical Imaging (ISBI)*, pages 1–5, 2024.

J. E. East, J. L. Vleugels, P. Roelandt, P. Bhandari, R. Bisschops, E. Dekker, C. Hassan, G. Horgan, R. Kiesslich, G. Longcroft-Wheaton, A. Wilson, and J.-M. Dumonceau. Advanced endoscopic imaging: European society of gastrointestinal endoscopy (ESGE) technology review. *Endoscopy*, 48(11):1029–1045, 2016.

D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems (NeurIPS)*, page 2366–2374, 2014.

R. Elvira, J. D. Tardós, and J. M. M. Montiel. CudaSIFT-SLAM: multiple-map visual SLAM for full procedure mapping in real human endoscopy. *arXiv preprint arXiv:2405.16932*, 2024.

J. Engel, T. Schöps, and D. Cremers. LSD-SLAM: Large-scale direct monocular SLAM. In *European Conf. on Computer Vision (ECCV)*, pages 834–849, 2014. ISBN 978-3-319-10605-2.

J. Engel, V. Usenko, and D. Cremers. A photometrically calibrated benchmark for monocular visual odometry. *arXiv preprint arXiv:1607.02555*, 2016.

J. Engel, V. Koltun, and D. Cremers. Direct sparse odometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(3):611–625, 2017.

J. M. Fácil, B. Ummenhofer, H. Zhou, L. Montesano, T. Brox, and J. Civera. CAM-Convs: Camera-aware multi-scale convolutions for single-view depth. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 11826–11835, 2019.

R. Fan, H. Wang, B. Xue, H. Huang, Y. Wang, M. Liu, and I. Pitas. Three-filters-to-normal: An accurate and ultrafast surface normal estimator. *IEEE Robotics and Automation Letters*, 6(3):5405–5412, 2021.

A. Fernandes-Araujo and J. M. M. Montiel. *Estimación de escala absoluta para Structure from Motion en endoscopio con fuente de luz cercana.* BSc thesis, EINA, Universidad de Zaragoza, 2022. URL `https://zaguan.unizar.es/record/120688`.

D. Freedman, Y. Blau, L. Katzir, A. Aides, I. Shimshoni, D. Veikherman, T. Golany, A. Gordon, G. Corrado, Y. Matias, and E. Rivlin. Detecting deficient coverage in colonoscopies. *IEEE Transactions on Medical Imaging*, 39(11):3451–3462, 2020.

H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao. Deep ordinal regression network for monocular depth estimation. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2002–2011, 2018.

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

C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow. Digging into self-supervised monocular depth estimation. In *IEEE Int. Conf. on Computer Vision (ICCV)*, pages 3828–3838, 2019.

J. J. Gómez-Rodríguez, J. Lamarca, J. Morlana, J. D. Tardós, and J. M. M. Montiel. SD-DefSLAM: Semi-direct monocular SLAM for deformable and intracorporeal scenes. In *IEEE Int. Conf. on Robotics and Automation (ICRA)*, pages 5170–5177, 2021.

J. J. Gómez-Rodríguez, J. M. M. Montiel, and J. D. Tardós. NR-SLAM: Non-rigid monocular SLAM. *IEEE Transactions on Robotics*, 40:4252–4264, 2024.

N. Gonçalves, D. Roxo, J. Barreto, and P. Rodrigues. Perspective shape from shading for wide-FOV near-lighting endoscopes. *Neurocomputing*, 150:136–146, 2015.

A. Gordon, H. Li, R. Jonschkowski, and A. Angelova. Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In *IEEE Int. Conf. on Computer Vision (ICCV)*, pages 8977–8986, 2019.

Y. Hao, J. Li, F. Meng, P. Zhang, G. Ciuti, P. Dario, and Q. Huang. Photometric stereo-based depth map reconstruction for monocular capsule endoscopy. *Sensors*, 20(18):5403, 2020a.

Y. Hao, M. Visentini-Scarzanella, J. Li, P. Zhang, G. Ciuti, P. Dario, and Q. Huang. Light source position calibration method for photometric stereo in capsule endoscopy. *Advanced Robotics*, 34(12):789–801, 2020b.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016.

C. Heckman, J. Morrison, and The Vicalib Team. Vicalib: Visual-inertial calibration tool. https://github.com/arpg/vicalib, 2016. University of Colorado, Boulder. (last accessed: July 15, 2022).

S. Hinterstoisser, S. Holzer, C. Cagniart, S. Ilic, K. Konolige, N. Navab, and V. Lepetit. Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes. In *IEEE Int. Conf. on Computer Vision (ICCV)*, pages 858–865, 2011.

D. Hong, W. Tavanapong, J. Wong, J. Oh, and P. C. De Groen. 3D reconstruction of virtual colon structures from colonoscopy images. *Computerized Medical Imaging and Graphics*, 38(1):22–33, 2014.

B. K. Horn and M. J. Brooks, editors. *Shape from Shading*. MIT Press, 1989.

B. Huang, J.-Q. Zheng, A. Nguyen, D. Tuch, K. Vyas, S. Giannarou, and D. S. Elson. Self-supervised generative adversarial network for depth estimation in laparoscopic images. In *Int. Conf. on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 227—-237. Springer, 2021.

S.-J. Hwang, S.-J. Park, G.-M. Kim, and J.-H. Baek. Unsupervised monocular depth estimation for colonoscope system using feedback network. *Sensors*, 21(8), 2021.

K. İncetan, I. O. Celik, A. Obeid, G. I. Gokceler, K. B. Ozyoruk, Y. Almalioglu, R. J. Chen, F. Mahmood, H. Gilbert, N. J. Durr, and M. Turan. VR-Caps: a virtual environment for capsule endoscopy. *Medical Image Analysis*, 70:101990, 2021.

R. Iranzo. *Estimación de Escala Absoluta a partir de Secuencias Monoculares de Colonoscopia*. BSc thesis, EINA, Universidad de Zaragoza, 2023.

R. Iranzo*, V. M. Batlle*, J. D. Tardós, and J. M. Montiel. EndoMetric: Near-light metric scale monocular SLAM. In *Int. Conf. on Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Springer, 2025.

H. Itoh, H. R. Roth, L. Lu, M. Oda, M. Misawa, Y. Mori, S.-e. Kudo, and K. Mori. Towards automated colonoscopy diagnosis: binary polyp size estimation via unsupervised depth learning. In *Int. Conf. on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 611–619. Springer, 2018.

H. Itoh, M. Oda, K. Jiang, Y. Mori, M. Misawa, S.-E. Kudo, K. Imai, S. Ito, K. Hotta, and K. Mori. Binary polyp-size classification based on deep-learned spatial information. *Int. J. of Computer Assisted Radiology and Surgery*, 16(10):1817–1828, 2021.

Y. Iwahori, H. Sugie, and N. Ishii. Reconstructing shape from shading images under point light source illumination. In *IEEE Int. Conference on Pattern Recognition*, volume 1, pages 83–87, 1990.

Y. Iwahori, S. Emoto, K. Funahashi, M. K. Bhuyan, A. Wang, and K. Kasugai. Recovering shape and size from a single endoscope image using optimization. In *Int. Congress on Advanced Applied Informatics (IIAI-AAI)*, pages 331–334, 2022.

S. Izquierdo and J. Civera. SfM-TTR: Using structure from motion for test-time refinement of single-view depth networks. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 21466–21476, 2023.

A. Johnston and G. Carneiro. Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 4756–4765, 2020.

J. T. Kajiya and B. P. Von Herzen. Ray tracing volume densities. *SIGGRAPH Comput. Graph.*, 18(3):165–174, 1984.

J. Kannala and S. S. Brandt. A generic camera model and calibration method for conventional, wide-angle, and fish-eye lenses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(8):1335–1340, 2006.

M. A. Karaoglu, N. Brasch, M. Stollenga, W. Wein, N. Navab, F. Tombari, and A. Ladikos. Adversarial domain feature adaptation for bronchoscopic depth estimation. In *Int. Conf. on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 300–310. Springer, 2021.

A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollar, and R. Girshick. Segment anything. In *IEEE Int. Conf. on Computer Vision (ICCV)*, pages 4015–4026, 2023.

G. Klein and D. Murray. Parallel tracking and mapping for small AR workspaces. In *IEEE and ACM International Symposium on Mixed and Augmented Reality*, pages 225–234, 2007.

I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab. Deeper depth prediction with fully convolutional residual networks. In *IEEE Int. Conf. on 3D Vision (3DV)*, pages 239–248, 2016.

J. Lamarca, S. Parashar, A. Bartoli, and J. M. M. Montiel. DefSLAM: Tracking and mapping of deforming scenes from monocular sequences. *IEEE Transactions on Robotics*, 37(1):291–303, 2020.

J.-H. Lee, M. Heo, K.-R. Kim, and C.-S. Kim. Single-image depth estimation based on fourier domain analysis. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 330–339, 2018.

L. Lettry, K. Vanhoey, and L. Van Gool. Unsupervised deep single-image intrinsic decomposition using illumination-varying image sequences. In *Computer Graphics Forum*, volume 37, pages 409–419, 2018.

Y. Li, C. Xie, H. Lu, X. Chen, J. Xiao, and H. Zhang. Scale-aware monocular SLAM based on convolutional neural network. In *IEEE Int. on Information Automation (ICIA)*, pages 51–56, 2018a.

Z. Li, Z. Xu, R. Ramamoorthi, K. Sunkavalli, and M. Chandraker. Learning to reconstruct shape and spatially-varying reflectance from a single image. *ACM Transactions on Graphics*, 37(6), 2018b. ISSN 0730-0301.

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 *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2475–2484, 2020.

Z. Li, X. Wang, X. Liu, and J. Jiang. BinsFormer: Revisiting adaptive bins for monocular depth estimation. *IEEE Transactions on Image Processing*, 33:3964–3976, 2024.

D. Lichy, J. Wu, S. Sengupta, and D. W. Jacobs. Shape and material capture at home. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 6123–6133, 2021.

F. Liu, C. Shen, G. Lin, and I. Reid. Learning depth from single monocular images using deep convolutional neural fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(10):2024–2039, 2015.

X. Liu, A. Sinha, M. Ishii, G. D. Hager, A. Reiter, R. H. Taylor, and M. Unberath. Dense depth estimation in monocular endoscopy with self-supervised learning methods. *IEEE Transactions on Medical Imaging*, 39(5):1438–1447, 2019.

X. Liu, Z. Li, M. Ishii, G. D. Hager, R. H. Taylor, and M. Unberath. SAGE: SLAM with appearance and geometry prior for endoscopy. In *IEEE Int. Conf. on Robotics and Automation (ICRA)*, pages 5587–5593, 2022.

H. Luo, Q. Hu, and F. Jia. Details preserved unsupervised depth estimation by fusing traditional stereo knowledge from laparoscopic images. *Healthcare Technology Letters*, 6(6):154, 2019.

X. Luo, J.-B. Huang, R. Szeliski, K. Matzen, and J. Kopf. Consistent video depth estimation. *ACM Transactions on Graphics*, 39(4):71–1, 2020.

R. Ma, R. Wang, S. Pizer, J. Rosenman, S. K. McGill, and J.-M. Frahm. Real-time 3D reconstruction of colonoscopic surfaces for determining missing regions. In *Int. Conf. on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 573–582. Springer, 2019.

R. Ma, R. Wang, Y. Zhang, S. Pizer, S. K. McGill, J. Rosenman, and J.-M. Frahm. RNNSLAM: Reconstructing the 3D colon to visualize missing regions during a colonoscopy. *Medical Image Analysis*, 72:102100, 2021.

F. Mahmood and N. J. Durr. Deep learning and conditional random fields-based depth estimation and topographical reconstruction from conventional endoscopy. *Medical Image Analysis*, 48:230–243, 2018.

F. Mahmood, R. Chen, and N. J. Durr. Unsupervised reverse domain adaptation for synthetic medical images via adversarial training. *IEEE Transactions on Medical Imaging*, 37(12):2572–2581, 2018.

N. Mahmoud, T. Collins, A. Hostettler, L. Soler, C. Doignon, and J. M. M. Montiel. Live tracking and dense reconstruction for handheld monocular endoscopy. *IEEE Transactions on Medical Imaging*, 38(1):79–89, 2018.

S. Mathew, S. Nadeem, S. Kumari, and A. Kaufman. Augmenting colonoscopy using extended and directional CycleGAN for lossy image translation. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 4696–4705, 2020.

H. Matsuki, R. Murai, P. H. Kelly, and A. J. Davison. Gaussian splatting SLAM. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 18039–18048, 2024.

R. Mecca, A. Wetzler, A. M. Bruckstein, and R. Kimmel. Near field photometric stereo with point light sources. *SIAM Journal on Imaging Sciences*, 7(4):2732–2770, 2014.

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. *Communications of the ACM*, 65(1):99–106, 2021.

R. Modrzejewski, T. Collins, A. Hostettler, J. Marescaux, and A. Bartoli. Light modelling and calibration in laparoscopy. *Int. J. of Computer Assisted Radiology and Surgery*, 15(5):859–866, 2020.

R. Mur-Artal and J. D. Tardós. ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras. *IEEE Transactions on Robotics*, 33(5): 1255–1262, 2017.

R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós. ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015.

R. A. Newcombe, S. J. Lovegrove, and A. J. Davison. DTAM: Dense tracking and mapping in real-time. In *IEEE Int. Conf. on Computer Vision (ICCV)*, pages 2320–2327, 2011.

T. Okatani and K. Deguchi. Shape reconstruction from an endoscope image by shape from shading technique for a point light source at the projection center. *Computer Vision and Image Understanding*, 66(2):119–131, 1997.

K. B. Ozyoruk, G. I. Gokceler, T. L. Bobrow, G. Coskun, K. Incetan, Y. Almalioglu, F. Mahmood, E. Curto, L. Perdigoto, M. Oliveira, S. Hasan, H. Araujo, H. Alexandrino, N. J. Durr, H. B. Gilbert, and M. Turan. EndoSLAM dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos. *Medical Image Analysis*, 71:102058, 2021.

K. Park, U. Sinha, J. T. Barron, S. Bouaziz, D. B. Goldman, S. M. Seitz, and R. Martin-Brualla. Nerfies: Deformable neural radiance fields. In *IEEE Int. Conf. on Computer Vision (ICCV)*, pages 5865–5874, 2021.

V. Parot, D. Lim, G. Gonzalez, G. Traverso, N. S. Nishioka, B. J. Vakoc, and N. J. Durr. Photometric stereo endoscopy. *Journal of Biomedical Optics*, 18(7):076017, 2013.

V. Patil, C. Sakaridis, A. Liniger, and L. Van Gool. P3Depth: Monocular depth estimation with a piecewise planarity prior. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1610–1621, 2022.

M. Poggi, F. Aleotti, F. Tosi, and S. Mattoccia. On the uncertainty of self-supervised monocular depth estimation. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3227–3237, 2020.

E. Prados and O. Faugeras. Shape from shading: a well-posed problem? In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 870–877, 2005.

Y. Quéau, T. Wu, and D. Cremers. Semi-calibrated near-light photometric stereo. In *Int. Conf. on Scale Space and Variational Methods in Computer Vision (SSVM)*, pages 656–668. Springer, 2017.

R. Ranftl, A. Bochkovskiy, and V. Koltun. Vision transformers for dense prediction. In *IEEE Int. Conf. on Computer Vision (ICCV)*, pages 12179–12188, 2021.

A. Rau, P. E. Edwards, O. F. Ahmad, P. Riordan, M. Janatka, L. B. Lovat, and D. Stoyanov. Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy. *Int. J. of Computer Assisted Radiology and Surgery*, pages 1–10, 2019.

A. Rau, B. Bhattarai, L. Agapito, and D. Stoyanov. Bimodal camera pose prediction for endoscopy. *IEEE Transactions on Medical Robotics and Bionics*, 5(4):978–989, 2023.

D. Recasens, J. Lamarca, J. M. Fácil, J. M. M. Montiel, and J. Civera. Endo-Depth-and-Motion: Reconstruction and tracking in endoscopic videos using depth networks and photometric constraints. *IEEE Robotics and Automation Letters*, 6(4):7225–7232, 2021.

J. Rodríguez-Puigvert, R. Martínez-Cantín, and J. Civera. Bayesian deep neural networks for supervised learning of single-view depth. *IEEE Robotics and Automation Letters*, 7(2):2565–2572, 2022.

J. Rodriguez-Puigvert, D. Recasens, J. Civera, and R. Martinez-Cantin. On the uncertain single-view depths in colonoscopies. In *Int. Conf. on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 130–140. Springer, 2022.

J. Rodríguez-Puigvert*, V. M. Batlle*, J. M. M. Montiel, R. Martinez-Cantin, P. Fua, J. D. Tardós, and J. Civera. LightDepth: Single-view depth self-supervision from illumination decline. In *IEEE Int. Conf. on Computer Vision (ICCV)*, pages 21273–21283, 2023.

J. Rodríguez-Puigvert, V. M. Batlle, J. D. Tardós, J. M. M. Montiel, J. Civera, R. Martinez-Cantin, and P. Fua. Self-supervised method for obtaining depth, albedo and surface orientation estimates of a space illuminated by a light source, 2023. Patent application PCT/EP2024/066877.

O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Int. Conf. on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 234–241. Springer, 2015.

D. Rückert, Y. Wang, R. Li, R. Idoughi, and W. Heidrich. Neat: Neural adaptive tomography. *ACM Transactions on Graphics*, 41(4), 2022.

S. Sang and M. Chandraker. Single-shot neural relighting and SVBRDF estimation. In *European Conf. on Computer Vision (ECCV)*, pages 85–101. Springer, 2020.

A. Saxena, S. Chung, and A. Ng. Learning depth from single monocular images. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 18, 2005.

D. Scaramuzza, A. Martinelli, and R. Siegwart. A toolbox for easily calibrating omnidirectional cameras. In *IEEE Int. Conf. on Intelligent Robots and Systems (IROS)*, pages 5695–5701, 2006.

A. Schmidt, O. Mohareri, S. DiMaio, M. C. Yip, and S. E. Salcudean. Tracking and mapping in medical computer vision: A review. *Medical Image Analysis*, 94, 2024. 103131.

J. L. Schönberger and J.-M. Frahm. Structure-from-Motion revisited. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 4104–4113, 2016.

J. L. Schönberger, E. Zheng, J.-M. Frahm, and M. Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European Conf. on Computer Vision (ECCV)*, pages 501–518, 2016.

A. Sengupta and A. Bartoli. Colonoscopic 3D reconstruction by tubular non-rigid structure-from-motion. *Int. J. of Computer Assisted Radiology and Surgery*, 16(7): 1237–1241, 2021.

S. Sengupta, A. Kanazawa, C. D. Castillo, and D. W. Jacobs. SfSNet: Learning shape, refectance and illuminance of faces in the wild. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.

J. Shan, Y. Li, T. Xie, and H. Wang. ENeRF-SLAM: a dense endoscopic SLAM with neural implicit representation. *IEEE Transactions on Medical Robotics and Bionics*, 6(3):1030–1041, 2024.

L. Sharan, L. Burger, G. Kostiuchik, I. Wolf, M. Karck, R. De Simone, and S. Engelhardt. Domain gap in adapting self-supervised depth estimation methods for stereo-endoscopy. *Current Directions in Biomedical Engineering*, 6(1), 2020.

M. Shen, Y. Gu, N. Liu, and G.-Z. Yang. Context-aware depth and pose estimation for bronchoscopic navigation. *IEEE Robotics and Automation Letters*, 4(2), 2019.

C. Shu, K. Yu, Z. Duan, and K. Yang. Feature-metric loss for self-supervised learning of depth and egomotion. In *European Conf. on Computer Vision (ECCV)*, pages 572–588. Springer, 2020.

P. P. Srinivasan, T. Wang, A. Sreelal, R. Ramamoorthi, and R. Ng. Learning to synthesize a 4D RGBD light field from a single image. In *IEEE Int. Conf. on Computer Vision (ICCV)*, 2017.

H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71 (3):209–249, 2021.

V. B. Surya Prasath and H. Kawanaka. Near-light perspective shape from shading for 3D visualizations in endoscopy systems. In *IEEE Int. Conf. on Bioinformatics and Biomedicine (BIBM)*, pages 2293–2295, 2017.

R. Szeliski. *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010.

Z. Teed and J. Deng. DROID-SLAM: Deep visual SLAM for monocular, stereo, and RGB-D cameras. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:16558–16569, 2021.

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*, volume 41, pages 703–735, 2022.

L. Tiwari, P. Ji, Q.-H. Tran, B. Zhuang, S. Anand, and M. Chandraker. Pseudo RGB-D for self-improving monocular SLAM and depth prediction. In *European Conf. on Computer Vision (ECCV)*, pages 437–455, 2020.

H. N. Tokgozoglu, E. M. Meisner, M. Kazhdan, and G. D. Hager. Color-based hybrid reconstruction for endoscopy. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) Workshop*, pages 8–15, 2012.

S. Tsuda, Y. Iwahori, M. K. Bhuyan, R. J. Woodham, and K. Kasugai. Recovering 3D shape with absolute size from endoscope images using RBF neural network. *Journal of Biomedical Imaging*, 2015, 2015. ISSN 1687-4188.

M. Turan, E. P. Ornek, N. Ibrahimli, C. Giracoglu, Y. Almalioglu, M. F. Yanik, and M. Sitti. Unsupervised odometry and depth learning for endoscopic capsule robots. In *IEEE Int. Conf. on Intelligent Robots and Systems (IROS)*, 2018.

T. van Dijk and G. de Croon. How do neural networks see depth in single images? In *IEEE Int. Conf. on Computer Vision (ICCV)*, pages 2183–2191, 2019.

M. Visentini-Scarzanella and H. Kawasaki. Simultaneous camera, light position and radiant intensity distribution calibration. In *Image and Video Technology*, pages 557–571, 2015.

M. Visentini-Scarzanella, D. Stoyanov, and G.-Z. Yang. Metric depth recovery from monocular images using shape-from-shading and specularities. In *IEEE Int. Conf. on Image Processing (ICIP)*, pages 25–28, 2012.

M. Visentini-Scarzanella, T. Sugiura, T. Kaneko, and S. Koto. Deep monocular 3D reconstruction for assisted navigation in bronchoscopy. *Int. J. of Computer Assisted Radiology and Surgery*, 12:1089–1099, 2017.

D. von Renteln, R. Djinbachian, M. Zarandi-Nowroozi, and M. Taghiakbari. Measuring size of smaller colorectal polyps using a virtual scale function during endoscopies. *Gut*, 72(3):417–420, 2023.

J. Wang, Y. Li, B. Chen, D. Cheng, F. Liao, T. Tan, Q. Xu, Z. Liu, Y. Huang, C. Zhu, W. Cao, L. Yao, Z. Wu, L. Wu, C. Zhang, B. Xiao, M. Xu, J. Liu, S. Li, and H. Yu. A real-time deep learning-based system for colorectal polyp size estimation by white-light endoscopy: development and multicenter prospective validation. *Endoscopy*, 56 (04):260–270, 2024a.

K. Wang, C. Yang, Y. Wang, S. Li, Y. Wang, Q. Dou, X. Yang, and W. Shen. EndoGSLAM: Real-time dense reconstruction and tracking in endoscopic surgeries using gaussian splatting. In *Int. Conf. on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 219–229. Springer, 2024b.

N.-H. Wang, R. Wang, Y.-L. Liu, Y.-H. Huang, Y.-L. Chang, C.-P. Chen, and K. Jou. Bridging unsupervised and supervised depth from focus via all-in-focus supervision. In *IEEE Int. Conf. on Computer Vision (ICCV)*, pages 12621–12631, 2021a.

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 (NeurIPS)*, volume 34, pages 27171–27183, 2021b.

Y. Wang, Y. Long, S. H. Fan, and Q. Dou. Neural rendering for stereo 3D reconstruction of deformable tissues in robotic surgery. In *Int. Conf. on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 431–441. Springer, 2022.

Y. Wang, Q. Han, M. Habermann, K. Daniilidis, C. Theobalt, and L. Liu. NeuS2: Fast learning of neural implicit surfaces for multi-view reconstruction. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3295–3306, 2023.

J. Watson, O. Mac Aodha, V. Prisacariu, G. Brostow, and M. Firman. The temporal opportunist: Self-supervised multi-frame monocular depth. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1164–1174, 2021.

R. J. Woodham. Photometric method for determining surface orientation from multiple images. *Optical engineering*, 19(1):139–144, 1980.

C. Wu, S. G. Narasimhan, and B. Jaramaz. A multi-image shape-from-shading framework for near-lighting perspective endoscopes. *International Journal of Computer Vision*, 86(2):211–228, 2010.

K. Xian, J. Zhang, O. Wang, L. Mai, Z. Lin, and Z. Cao. Structure-guided ranking loss for single image depth prediction. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 608–617, 2020.

K. Xu, Z. Chen, and F. Jia. Unsupervised binocular depth prediction network for laparoscopic surgery. *Computer Assisted Surgery*, 24(sup1):30–35, 2019.

J. Yan, H. Zhao, P. Bu, and Y. Jin. Channel-wise attention-based network for self-supervised monocular depth estimation. In *IEEE Int. Conf. on 3D Vision (3DV)*, pages 464–473, 2021.

Q. Yan, P. Ji, N. Bansal, Y. Ma, Y. Tian, and Y. Xu. Fisheyedistill: Self-supervised monocular depth estimation with ordinal distillation for fisheye cameras. *arXiv preprint arXiv:2205.02930*, 2022.

N. Yang, L. v. Stumberg, R. Wang, and D. Cremers. D3VO: Deep depth, deep pose and deep uncertainty for monocular visual odometry. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1281–1292, 2020.

Z. Yang, P. Wang, W. Xu, L. Zhao, and R. Nevatia. Unsupervised learning of geometry from videos with edge-aware depth-normal consistency. In *Thirty-Second AAAI Conference on Artificial Intelligence*. AAAI Press, 2018. ISBN 978-1-57735-800-8.

M. Yoshioka, Y. Sakaguchi, D. Utsunomiya, S. Sonoda, T. Tatsuta, S. Ozawa, Y. Teramura, K. Harada, H. Kinugasa, and H. Okada. Virtual scale function of gastrointestinal endoscopy for accurate polyp size estimation in real-time: a preliminary study. *Journal of Biomedical Optics*, 26(9):096002–096002, 2021.

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

S. Zhang, L. Zhao, S. Huang, M. Ye, and Q. Hao. A template-based 3D reconstruction of colon structures and textures from stereo colonoscopic images. *IEEE Transactions on Medical Robotics and Bionics*, 3(1):85–95, 2020.

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. *ACM Transactions on Graphics*, 40(6):1–18, 2021.

Y. Zhang, J. Sun, X. He, H. Fu, R. Jia, and X. Zhou. Modeling indirect illumination for inverse rendering. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 18643–18652, 2022.

Q. Zhao, T. Price, S. Pizer, M. Niethammer, R. Alterovitz, and J. Rosenman. The endoscopogram: A 3D model reconstructed from endoscopic video frames. In *Int. Conf. on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 439–447. Springer, 2016.

H. Zhou, B. Ummenhofer, and T. Brox. DeepTAM: Deep tracking and mapping. In *European Conf. on Computer Vision (ECCV)*, pages 851–868. Springer, 2018a.

J. Zhou, A. Das, F. Li, and B. Li. Circular generalized cylinder fitting for 3D reconstruction in endoscopic imaging based on MRF. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) Workshop*, pages 1–8, 2008.

Q.-Y. Zhou, J. Park, and V. Koltun. Open3D: A modern library for 3D data processing. *arXiv preprint arXiv:1801.09847*, 2018b.

T. Zhou, M. Brown, N. Snavely, and D. G. Lowe. Unsupervised learning of depth and ego-motion from video. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1851–1858, 2017.

Z. Zhou, X. Fan, P. Shi, and Y. Xin. R-MSFM: Recurrent multi-scale feature modulation for monocular depth estimating. In *IEEE Int. Conf. on Computer Vision (ICCV)*, pages 12777–12786, 2021.

J. Zubizarreta, I. Aguinaga, and J. M. M. Montiel. Direct sparse mapping. *IEEE Transactions on Robotics*, 36(4):1363–1370, 2020.

This thesis was supported by EU-H2020 grant 863146: ENDOMAPPER, Spanish government grants FPU20/06782 and PID2021-127685NB-I00 and by Aragón government grants DGA T45-17R and T45-23R.