TESIS DE LA UNIVERSIDAD
DE ZARAGOZA

2024

164

Javier Rodríguez Puigvert

# Uncertainty and Self-Supervision in Single-View Depth

Director/es

Civera Sancho, Javier
Martínez Cantín, Rubén

ISSN 2254-7606

Prensas de la Universidad
Universidad Zaragoza

Universidad de Zaragoza
Servicio de Publicaciones

ISSN 2254-7606

Tesis Doctoral

# UNCERTAINTY AND SELF-SUPERVISION IN SINGLE-VIEW DEPTH

Autor

Javier Rodríguez Puigvert

Director/es

Civera Sancho, Javier
Martínez Cantín, Rubén

**UNIVERSIDAD DE ZARAGOZA**
**Escuela de Doctorado**

Programa de Doctorado en Ingeniería de Sistemas e Informática

2023

Universidad Zaragoza

1542

# Tesis Doctoral

## Uncertainty and Self-Supervision in Single-View Depth

Autor

Javier Rodríguez Puigvert

Directores

Javier Civera Sancho

Rubén Martínez Cantín

ESCUELA DE DOCTORADO
2023

*Für Nina, meine Frau.*
*A Carolina y Nuria, mis hermanas.*
*A Nuria y Alfonso, mis padres.*

# Agradecimientos

Hace cuatro años, decidí embarcarme en un camino apasionante, lleno de desafíos, dejando atrás una vida acomodada en Alemania, pero llevándome lo mejor de ella. Mi mujer, Nina, muchas gracias por estar a mi lado y apoyarme incondicionalmente. Gracias por atreverte a tomar la valiente decisión de trasladarnos a Zaragoza y emprender esta aventura, *fortis fortuna iuvat*.

Me gustaría expresar mi gratitud más sincera a mis directores Javier y Rubén, por motivarme y guiarme durante estos años de mi tesis. Muchas gracias por vuestro tiempo y dedicación. Bajo vuestra supervisión, he aprendido las habilidades analíticas para desafiar y abordar problemas complejos. A ti Javier, muchas gracias por darme la oportunidad de hacer el doctorado bajo tu supervisión, sin conocerme demasiado. Todavía me acuerdo de la primera conversación que tuvimos en el ISMAR de Munich, hace ya 5 años y la trascendencia que tuvo en mi vida.

Quisiera agradecer a Pascal, Mingo y Monti, por darme siempre un feedback sincero y constructivo. Trabajando con vosotros, me habéis trasladado la importancia de tener un pensamiento crítico y positivo sobre la investigación. Quisiera agradecer también a César que me diera la oportunidad de trabajar junto a él durante mi estancia en la ETH Zurich. Muchas gracias a todos los compañeros de unizar y del L.108, en especial a, Juanjo, Sergio, Julio, Richard, David, León, Edu, Lorenzo, Tomás, Morlana, Victor y Bruno. Nuestras conversaciones siempre han sido super enriquecedoras, siempre he aprendido algo nuevo con vosotros.

Gracias mis amigos Tono, por ayudarme con cosas de Unity, Alberto, por ayudarme distintas discusiones y Arnau por ayudarme con su conocimiento de programación web. A David y Luis Miguel, por estar ahí desde 2008 cuando empezábamos nuestro camino para ser ingenieros.

Muchas gracias a mi Familia, Rodríguez, Puigvert y Hoppmann, por su amor y apoyo. Especialmente a mis hermanas Nuria y Carolina que siempre han sido un ejemplo para mí.

Finalmente, este doctorado está dedicado a mis padres Nuria y Alfonso. Mi padre, Alfonso Rodríguez Cid, la persona más perseverante que he conocido, inteligente y perspicaz donde los haya. Siento tu apoyo diario, aunque no estés con nosotros.

# Resumen

La estimación de profundidad a partir de una sola vista se refiere a la capacidad de obtener información tridimensional por píxel a partir de una imagen RGB bidimensional. La profundidad a partir de una sola vista tiene gran relevancia en una amplia gama de aplicaciones en campos como la realidad aumentada, la realidad virtual, la robótica, la medicina o la conducción autónoma. Desde un punto de vista geométrico, la reconstrucción 3D a partir de una imagen es un problema mal condicionado, porque existen múltiples soluciones de profundidad que explican una imagen 2D. Recientemente, las redes neuronales profundas han demostrado ser muy eficaces para aprender los patrones de apariencia visual que corresponden con ciertas estructuras 3D. Sin embargo, la mayoría de los métodos son deterministas y no estiman la incertidumbre asociada a sus predicciones. Esto puede tener consecuencias desastrosas cuando se aplica a campos como la conducción autónoma o la robótica médica. En esta tesis hemos abordado este problema cuantificando la incertidumbre de la profundidad de una sola vista para redes neuronales profundas bayesianas.

La estimación de la profundidad en una sola vista puede ser muy eficaz cuando se dispone de suficientes datos de profundidad anotados para el entrenamiento supervisado. Sin embargo, existen escenarios, por ejemplo nuestra aplicación con imágenes endoscópicas, en los que no es posible obtener dichos datos. La estimación de profundidad a partir de imágenes endoscópicas es un requisito para una amplia gama de tecnologías de asistencia al personal médico basadas en IA, como la localización y medición precisas de tumores o la identificación de zonas no inspeccionadas. En esta tesis, presentamos un método que facilita la transición de los métodos entrenados en datos sintéticos al dominio real teniendo en cuenta las incertidumbres asociadas. En concreto, introducimos una arquitectura maestro-estudiante (teacher-student en inglés) consciente de la incertidumbre que se entrena de forma autosupervisada, teniendo en cuenta la incertidumbre del maestro entrenado en un conjunto de datos sintéticos.

Finalmente, en algunos dispositivos médicos como los endoscopios, la cámara y las fuentes de luz están rígidamente unidas y situadas a una pequeña distancia de las superficies a explorar. Como última contribución de esta tesis doctoral, modelamos el hecho de que para cualquier albedo y superficie dados, el brillo del píxel es inversamente proporcional al cuadrado de la distancia a la superficie y proponemos el uso de la iluminación como una potente señal de autosupervisión de una sola vista para redes neuronales profundas.

# Acknowledgements

Four years ago, I decided to embark on an exciting and challenging journey, leaving behind a comfortable life in Germany, but taking the best of it with me. My wife, Nina, thank you so much for being by my side and supporting me unconditionally. Thank you for bravely taking the courageous decision to move to Zaragoza and start this adventure (fortis fortuna iuvat). I would like to express my most sincere gratitude to my directors Javier and Rubén, for motivating and guiding me during these years of my thesis. Thank you very much for your time and dedication. Under your supervision, I have learned the analytical skills to challenge and tackle complex problems. I still remember the first conversation we had at ISMAR in Munich, 5 years ago now, and how significant it was in my life. I would like to thank Pascal, Mingo and Monti for always giving me honest and constructive feed-back. By working with you, you have shown me how important it is to think critically and positively about research. I would also like to thank César for giving me the opportunity to work with him during my stay at ETH Zurich. Many thanks to all my colleagues at unizar and L.108, especially Juanjo, Sergio, Julio, Richard, David, León, Edu, Lorenzo, Tomás, Morlana, Victor and Bruno. Our conversations have always been very enriching, I have always learned something new from you.

Thanks to my friends Tono, for helping me with Unity stuff, Alberto, for helping me with different discussions and Arnau for helping me with his programming knowledge. To David and Luis Miguel, for being there since 2008 when we were starting our way to become engineers.

Many thanks to my family, Rodríguez, Puigvert and Hoppmann, for their love and support. Especially to my sisters Nuria and Carolina who have always been an inspiration to me. Finally, this PhD is dedicated to my parents Nuria and Alfonso. My father, Alfonso Rodríguez Cid, is the most persevering person I have ever met. I feel your daily support, even though you are not with us.

# Abstract

Single-view depth estimation refers to the ability to derive three-dimensional information per pixel from a single two-dimensional RGB image. Estimating depth from a single image is required in a wide range of applications in fields such as augmented reality, virtual reality, robotics, medicine or autonomous driving.

Single-view depth estimation is an ill-posed problem because there are multiple depth solutions that explain 3D geometry from a single view. While deep neural networks have been shown to be effective at capturing depth from a single view, the majority of current methodologies are deterministic in nature. Accounting for uncertainty in the predictions can avoid disastrous consequences when applied to fields such as autonomous driving or medical robotics. We have addressed this problem by quantifying the uncertainty of supervised single-view depth for Bayesian deep neural networks. When there is enough ground truth depth data for supervised training, single-view depth estimation can be remarkably effective. There are scenarios, especially in medicine in the case of endoscopic images, where such annotated data is not available. Nevertheless, the estimation of depth information from endoscopic images is a prerequisite for a wide range of AI-based technologies, such as the accurate localisation and measurement of tumours or the identification of non-inspected areas.

To alleviate the lack of data, we present a method that improves the transition from synthetic to real domain methods. We introduce an uncertainty-aware teacher-student architecture that is trained in a self-supervised manner, taking into account the teacher uncertainty trained on a synthetic dataset.

Given the vast amount of unannotated data and the challenges associated with capturing annotated depth in medical minimally invasive procedures, we advocate a fully self-supervised approach that only requires RGB images and the geometric and photometric calibration of the endoscope. In endoscopic imaging, the camera and light sources are co-located at a small distance from the target surfaces. This setup indicates that brighter areas of the image are nearer to the camera, while darker areas are further away. Building on this observation, we exploit the fact that for any given albedo and surface orientation, pixel brightness is inversely proportional to the square of the distance. We propose the use of illumination as a strong single-view self-supervisory signal for deep neural networks.

# Index

# Chapter 1

# Introduction

Vision is tremendously powerful; we as humans are able to navigate through the world and most often it feels effortless, as our visual system has evolved through millennia to process information seamlessly. However, even if humans generally excel at making qualitative judgments about visual scenes, we are not so good on quantifying our perceptions. For example, we can easily qualify an object as near or far, but when it comes to an accurate quantitative measurement, such as estimating an object's distance in centimetres, we might be off by a gross margin. Looking at the more especific example in Figure 1.1, it is easy for us to understand the rough geometry and the semantic structure of the scene. We can make spatial sense of the space, inferring a back wall at the end of the two rows of posters, plan a path to get there and navigate through the scene, even if it is crowded. Such understanding is an extremely complex task, only made possible by our prior knowledge and our extraordinary pattern recognition abilities.

Computer vision aims to emulate, and hopefully surpass, human visual perception. In fact, computer vision's capabilities have already begun to outpace human abilities in specific areas [Kaufmann et al., 2023]. The field of computer vision is currently on a growth curve never seen before. From the early days of deep learning, when convolutional neural networks were used for image classification of some simple entities [LeCun et al., 2015], to recent advances in visual transformers [Dosovitskiy et al., 2020] that masterfully create features with a consistent global perspective across images or Neural Radiance Fields [Mildenhall et al., 2020] (NERFs), the field of computer vision is experiencing an extraordinary evolution. Even more recently, the introduction of diffusion models [Ho et al., 2020] has started yet another revolution in generative modelling and opened up new avenues in image synthesis and rendering.

As the capabilities of computer vision continue to expand and evolve, we expect it to have a disruptive effect on technologies such as vision-assisted robotics or medical image analysis. However, and despite the impressive recent advances, several computer vision tasks remain unsolved. Specifically and among others, single-view depth estimation

Figure 1.1: ICCV 2023 Poster Session

still presents significant challenges. Although depth can also be estimated from multiple views, single-view depth may be relevant when only one view of the scene is available. In addition, for deforming scenes, small camera motion or large illumination changes, multi-view reconstruction may present significant challenges and single-view depth be a reasonable alternative.

Single-view depth may be applied in a wide range use cases such as augmented reality [Luo et al., 2020], computational photography [Barron et al., 2015] or medical robotics [Kader et al., 2023]. This thesis addresses this specific topic, with particular focus on endoscopic imaging, and our work contains several contributions and experimental evaluations in this field with focus on self-supervision and uncertainty quantification.

## 1.1 Depth perception

Depth perception is a fundamental aspect of visual cognition. It refers to the ability to perceive the 3D spatial relationships of the objects within a scene, from the sole input of one or many 2D images. This cognitive capability is the basis for 3D reconstruction from a single or multiple views. The importance of depth perception covers a wide range of industries and applications.

In immersive technologies, depth perception is a must. For example, in augmented reality, it enables virtual objects to be placed in the real world with appropriate depth and hence coherence with the rest of the viewed scene. In virtual reality, users can experience a virtual environment while the device perceives the real environment and prevents collisions. In robotics and self-driving cars, depth perception is a cornerstone to enable full autonomy. 3D reconstruction plays a crucial role in the understanding of a robot's surroundings, as autonomous robots should ground their behaviours on a comprehensive model of their environment, whether navigating indoors or exploring outdoor terrain. 3D models estimated from onboard sensor data provide robots with detailed spatial information, enabling among others collision-free navigation, object recognition and safe interaction with the environment.

When focusing on 3D reconstruction, the task of single-view depth estimation has its own unique challenges and importance. It is critical in scenarios where there is frequent deformation, small baseline between frames, or the camera is frequently occluded. All these aspects are frequent in the medical arena, that has adopted vision technologies in many tasks. Nowadays, surgeons have the ability to visualise internal body structures in three dimensions, which allows for more precise planning and execution of surgical procedures. As another example, dentists also use 3D scanning to create accurate models for the design and manufacture of dental appliances.

Recovering depth from images has been studied extensively in the computer vision community. Multi-view, shape-from-X, and structure-from-motion are effective in recovering the three-dimensional structure of a scene using multiple images taken from different viewpoints. The estimation of depth from a single view is a potential complement to the many challenging conditions that may occur in multi-view setups, like low textured scenes, insufficient motion between views, deformations and illumination changes. While 3D reconstruction often leverages multiple views to recreate a scene or object in three dimensions, single-view depth estimation seeks to achieve a similar goal from only one perspective. However, it is an inherently complex problem due to its ambiguous nature. This ambiguity arises because multiple 3D configurations can result in the same 2D image, making it impossible from a general geometric perspective to identify the one that produce the image. This challenge has motivated considerable research into the use of additional cues or sophisticated algorithms to approximate depth more accurately. Despite its complexity, mastering single-view depth estimation is critical for scenarios where multi-view or motion-based methods are impractical or impossible, e.g., in endoscopy. As depth estimation algorithms from single views mature in precision, they also offer potential for integration into multi-view 3D reconstruction pipelines [Facil et al., 2017].

In this thesis, we tackle the problem of single-view depth from different perspectives, but always in a data-driven manner and deep neural networks. Firstly, we use data-driven methods and extract the uncertainty associated with depth maps. Secondly, we reduce the complexity of the problem by using illumination decline in the scenarios where the camera and illumination are co-located.

Deep neural network developments have swiftly progressed, becoming the predominant approach for various tasks within computer vision like semantic segmentation, classification object detection, or depth estimation. Specifically, CNNs [LeCun et al., 2015] can identify basic features in the early layers and more complex spatial associations in the deeper layers. CNNs apply a composition of convolutions to the input image, extracting hierarchically organized feature maps. The feature maps correspond to patterns in the input data such as corners, edges, textures, and in general local or mid-level patterns in the image.

A pioneering work in the recovery of 3D from 2D images was done by Saxena et al. [2005], who propose multi-spatial monocular clues that models the relative depths. Eigen et al. [2014] were the first ones to demonstrate an effective use of deep learning in this problem. They propose a supervised learning pipeline using a convolutional neural networks with a coarse-to-fine architecture, being the input a RGB image and the target per-pixel depth estimations. By improving the network architecture, Laina et al. [2016] proposed deeper fully convolutional models and later Fu et al. [2018] formulated the problem from an ordinal regression perspective in a discrete depth space.

Visual Transformers are the latest significant innovation in the field of deep learning. Based on self-attention mechanisms, visual transformers weigh the relevance of different image patches relative to each other. This mechanism helps to focus the attention on the most important areas, having a global context of the image. In contrary to the fundamentals of convolutional neural networks, visual patches are of fixed size and non-overlapping, that help to learn recognize patterns across different patches of the image. The architecture of transformers has been adapted to work with image patches visual transformers [Dosovitskiy et al., 2020] and later applied to a single-view depth estimation [Ranftl et al., 2021].

Annotated depth data is very diverse depending of the captured scene (indoors or outdoors), the nature of annotation (be it relative or absolute), and the accuracy (from lasers, synthetic data, or Structure from Motion). Ranftl et al. [2020] proposed the use of different source of depth using a scale and shifting invariant losses. In the medical domain, single-view depth estimation has been extensively studied for endoscopic purposes. Visentini-Scarzanella et al. [2017] used computed tomography renderings for depth supervision in bronchoscopies. However, computed tomography scans in particular and ground-truth depth annotations in general are very rare in endoscopy, which makes self-supervision methods essential for a practical application of this technology.

## 1.2 Self-supervised single-view depth

Capturing large volumes of annotated data is resource-intensive, often requires domain expertise and specific hardware. In areas such as medical imaging, where expert annotation is both essential and limited, the challenge is particularly pronounced. In endoscopy, for example, the size of the endoscopes can make it challenging to obtain such annotations. Self-supervised methods are emerging as an alternative to traditional ground-truth supervision, especially given the vast difference between the amount of data with and without annotations, the last one being several orders of magnitude larger.

In the field of single-view depth, seminal contributions have been made by Godard et al. [2017]. They proposed a self-supervised learning method that enforces multi-view photometric consistency using a left-right stereo camera. Similarly, Zhou et al. [2017] explored the use of monocular ego-motion to enforce photometric consistency with a monocular image, where the motion is predicted by a network. Godard et al. [2019] introduced a pipeline for multi-scale self-supervision using monocular, stereo and combined cameras. However, this type of supervision can introduce noise due to various factors such as inaccuracies in camera motion estimation, perspective distortions, occlusions, and non-Lambertian effects. In scenarios where obtaining real labeled data (or ground truth data) is challenging or it is not possible, there has been a tendency to generate synthetic data from simulators and apply learning techniques using synthetic data as reference. Even if the progress of realistic simulations is notable, the issue of domain change (from simulation to real) remains a persistent challenge. This domain change refers to the discrepancies between synthetic data generated by simulators and real-world data, which can hinder the performance of models trained only on simulated environments. The singularity of the medical domain has resulted in considerable research in the transition from synthetic to real-world scenarios. For example, the research by Shen et al. [2019] uses a conditional GAN for depth recovery, incorporating SLAM and multiview data. Karaoglu et al. [2021] focus on depth estimation from monocular images for broncoscopy navigation. The authors propose a domain adaptive pipeline in two steps: training a network in synthetic labels and use advesarial domain feature adaptation to enhance the performance on real images. In the work of Chen et al. [2019b], a depth network is trained with synthetic images of a simple colon model and fine-tuned with domain-randomized photo-realistic images rendered from computed tomography scans. In this thesis, we have proposed several contributions in the area of self-supervised learning for single-view depth. Firstly, in Chapter 3, we evaluate how uncertainties affect the domain change scenario and propose an uncertainty aware teacher-student architecture that mitigates the effect of domain change. Secondly, in Chapter 4, we introduce a new self-supervised learning technique that employs the information of the illumination decline as supervisory signal.

## 1.3 Single-view depth uncertainty

The aforementioned research on image-based depth detection uses deterministic deep learning models. This means they always produce the same output for a same input without a measure of how certain they are about the prediction. During training, a certain loss function is minimized in order to find the set of parameters that will give the best performance in new data. Such models do not take into account the inherent uncertainties in their predictions, and this is relevant for their use in real-world scenarios. These uncertainties can occur due to the distribution of the data, the lack of data or the inherent limitations of the models. Uncertainty allows us to assess the reliability of predictions, which can be used to guide subsequent actions, particularly in fields such as medicine and autonomous driving, where decisions can have fatal consequences. There are two types of uncertainty that can be considered in deep learning: aleatoric uncertainty and epistemic uncertainty [Kiureghian and Ditlevsen, 2009]. Aleatoric uncertainty is the uncertainty of the observations, and it is inherent to the data, i.e. it would still exist even if more data is available. This uncertainty can be thought of as noise in the sensor . For example, in the case of single-view depth, when RGB-D cameras are used, the camera depth sensor noise is modelled by the aleatoric uncertainty. Aleatoric uncertainty is further subdivided into homoscedastic uncertainty, that is independent of the inputs (constant), and heteroscedastic uncertainty that is dependent on the inputs to the model. Epistemic uncertainty refers to the uncertainty of the model, and consequently, Bayesian deep learning is a natural way to capture epistemic uncertainty in neural networks. In Bayesian deep learning, the epistemic uncertainty can be obtained using a prior distribution $p(\theta)$ over the neural network parameters $\theta$ and computing its posterior distribution $p(\theta|\mathcal{D})$ given a dataset $\mathcal{D}$ using Bayes rule: $p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}$. This equation is intractable for deep learning architectures. However, there are scalable approaches to deep learning such as MC dropout and deep ensembles. In [Kendall and Gal, 2017], a seminal effort is made to highlight the significant types of uncertainty in deep learning, namely aleatoric and epistemic uncertainties. Kendall and Gal [2017] combine epistemic and aleatoric uncertainty by using a MC dropout approximation of the posterior distribution. Modelling aleatoric uncertainty as observational noise through a network output and epistemic uncertainty as the variations in the prediction by several forward passes. Deep ensembles [Lakshminarayanan et al., 2017] consist of training the same architecture multiple times, assuming different random initialisation of its parameters. Each ensemble is considered to be trained to a local minimum. The set of ensembles are an approximation of the model distribution.

For single-view self-supervised learning, Poggi et al. [2020] propose to extract depth uncertainty of depth from labels of a pretrained self-supervised network by introducing a teacher-student architecture. It is important to note that uncertainty measures are of limited value unless they are calibrated with the corresponding errors. In this thesis, we explore deeply MC dropout and deep ensembles in a supervised single-view depth setting, evaluating both their relative and absolute calibration of uncertainty.

## 1.4 Endoscopic single-view depth estimation

Colonoscopy and gastroscopy are among the most common procedures performed in hospitals. During an endoscopy, a physician uses an endoscope to examine a patient. This procedure involves navigating through the patient's body without damaging tissues and identifying and treating areas that represent a health risk.

Navigating environments such as the colon is a complex task that largely depends on the expertise of the doctor and the preparation of the patient. The primary tool aiding this procedure is the endoscope, with the camera at its tip serving as the unique sensor. This navigation relies heavily on a doctor prior anatomical knowledge and the real-time visual and mechanical feedback from the endoscope. Despite these aids, there remains a risk of perforation, emphasizing the need for enhanced depth perception to assist clinicians during navigation.

Concretely, single-view depth estimation may potentially revolutionize endoscopic mapping and navigation. This method emphasizes creating 3D maps of the inspected areas, allowing for a more detailed understanding of the internal environment [Aides et al., 2020]. Given the predominance of monocular cameras in endoscopic procedures, deep learning algorithms present significant promise for advancing this technology. However, images from endoscopic procedures often contain challenges such as liquids, specular lighting, tissue deformations, drastic illumination changes, camera movements, and minimal parallax. Therefore, we argue in favor of systems based on single-view depth, which do not necessitate calculations of the camera relative position.

During screening procedures, physicians aim to identify polyps and tissue anomalies. Assessing these findings is crucial for timely and accurate intervention [Kader et al., 2023]. In this context, single-view depth can be instrumental in measuring polyps and in assessing areas affected by conditions like Crohn's disease.

Furthermore, the integration of single-view depth technology can significantly enhance robot-assisted, computer-aided interventions. By offering enhanced perception during navigation, single-view depth will benefit both robotic and manual procedures, potentially reducing associated risks to patients.

## 1.5 Our contribution to single-view depth:

The work described in this thesis presents significant improvements in the field of single-view depth estimation by using deep neural networks. The improvements encompass uncertainty quantification for single-view depth, a self-supervised approach based on a teacher-student architecture that models the teacher uncertainty, and a new single-view self-supervision method based on illumination decline.

Previous research have shown that deep neural networks have significant potential to make depth estimations, which gives a promising basis to further develop their usage. They offer an appearance-based approach to the ill-posed problem of depth perception from 2D images. In addition, deep ensembles and MC dropout have shown to be a scalable alternative to Bayesian deep learning methods.

The first of our contributions is the identification of deep ensembles as the best calibrated method for scalable Bayesian depth deep learning methods. We demonstrate empirically that adding dropout in all layers of the encoder delivers better results than other variations of MC dropout. This result is of practical relevance in the estimation of depth and uncertainty, because MC dropout requires much less memory than deep ensembles. It is particularly attractive for systems with limited resources due to its reduced memory requirements. As a result, MC dropout emerges as a potentially more adaptable approach when considering a Bayesian approach for real-world applications. We also show the application of Bayesian depth networks in the context of 3D reconstruction by applying pseudo-RGB Depth Iterative Closest Point (ICP), showing that relative transformation can be improved by excluding the points with higher uncertainties.

Supervised methods outperform self-supervised methods in terms of accuracy, primarily because the benefit from training with accurate and annotated data. There are situations where acquiring labeled data is difficult, being an inclination to use synthetic data from simulators as a reference for learning. However, the domain shift problem is still a challenge because of the difference between synthetic and real images. In medical settings, especially in procedures like colonoscopies, the challenges presented by domain change become especially pronounced.

We pioneer the exploration of uncertainty quantification for single-view depth specifically focusing on medical colonoscopy images, marking a novel contribution to the existing literature. Colonoscopy images are particularly difficult as they present many challenges. They have numerous discontinuities mainly due to partial occlusions in folds or haustra. The lighting conditions within the colon can vary significantly depending on the endoscope electronics, leading to frequent illumination changes. On top of that, the presence of specular light, which results from the reflection of light off the wet surfaces inside the colon, adds another layer of complexity to the image interpretation. Due to this conditions, colonoscopy images are an ideal benchmark for evaluating the effectiveness and reliability of single-view depth estimation methods. The importance of understanding uncertainty in this domain is crucial, as inaccurate or misleading predictions could lead to misdiagnoses or oversight of critical medical conditions. Therefore, it is critical to have a robust system that not only estimates depth, but also provides a measure of the uncertainty associated with that estimation. Our main research advance is the introduction of a self-supervised method anchored in an uncertainty-aware teacher-student architecture. Our results demonstrate that our approach excels other types of learning in terms of depth and uncertainty quantification. Moreover, the results show the advantages of weakly supervised learning with respect to self-supervised learning.

Until now, single-view depth has traditionally been classified in two categories: supervised learning, which relies on depth annotations or annotations derived from a multi-view system; and self-supervised learning, which leverages geometric and/or photometric consistency across multiple views. We introduce a new form of depth supervision derived from the illumination decline principle. The main result of our work is a unique single-view self-supervised learning method for depth estimation. This method is applicable in dark environments, where the light source and camera are co-located. Such conditions are prevalent in numerous real-world scenarios, including within the human body as seen in clinical procedures which involve examining the interior of hollow organs or cavities, underwater domains like in submarines or specific seabed regions, enclosed natural formations such as caves, rigid enclosures often encountered during engine inspections or examinations of other mechanical parts, and confined cylindrical spaces like pipes or other similar structures that are typically hard to access.

For the development of this method, we grounded our research within a medical context, specifically focusing on endoscopes including procedures like colonoscopies and gastroscopies.

The medical field, especially when it comes to gastroscopy, is full of challenges reminiscent of those described for colonoscopy. Gastroscopy images often show foam, varying textures, numerous specular reflections and a variety of shapes that differ from what we observe in other endoscopic procedures, particularly when examining regions such as the stomach or other related areas. Despite these complexities, this environment is a natural platform for our research. The particularity lies in endoscopic imaging, where the imaging device is the only light source illuminating the scene. The foundation of our approach is based on a simple yet profound observation: In endoscopic images, brighter areas tend to be closer to the camera, while darker areas tend to be further away. This insight differs significantly from conventional self-supervised learning paradigms. Our method is unique in that it requires only a single reference image during the learning process. Our technique not only outperforms other self-supervised methods, but also rivals depth-supervised strategies. In the scientific literature, self-supervised techniques have always lagged far behind supervised ones, being ours the first method that is competitive with ground truth supervision. An additional point in favour of our single-view self-supervision is its resilience to domain changes. Specifically, a distinct advantage of our method is the capability to refine the depth estimations on-the-fly at test time, which opens the way for improved predictions.

In addition to predicting depth, our technique is adept at assessing the albedo of the tissue. This advance is also significant, given the inherent value of this prediction mode. By predicting the albedo, our method can become instrumental in detecting flat lesions and texture variations, which are critical in identifying conditions such as Barrett's oesophagus in gastroscopy or Crohn's disease indications in colonoscopy. In particular, albedo predictions can improve the effectiveness of other machine learning systems. These diverse applications highlight the transformative potential of our method in medical diagnostics.

## 1.6 List of publications

Throughout this thesis, the following publications have been published (although the thesis document focuses in the first three):

- Rodríguez-Puigvert, J., Martínez-Cantín, R., Civera, J. (2022). **Bayesian deep neural networks for supervised learning of single-view depth.** IEEE Robotics and Automation Letters (R-AL), 7(2), 2565-2572. presented at IEEE International Conference on Robotics and Automation (ICRA) 2022

- Rodriguez-Puigvert, J., Recasens, D., Civera, J., Martinez-Cantin, R. (2022, September). **On the uncertain single-view depths in colonoscopies.** In International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) (pp. 130-140). Cham: Springer Nature Switzerland.

- Rodriguez-Puigvert, J., M. Battle, V., Montiel, J., Cantin, R., Fua, P., Tardos, J., Civera, J. (2023). **LightDepth: Single-View Depth Self-Supervision from Illumination Decline.** In International Conference on Computer Vision (ICCV) 2023.

- Patent Pending: EP23382614.8. **SELF-SUPERVISED METHOD FOR OBTAINING DEPTH, ALBEDO AND SURFACE ORIENTATION ESTIMATES OF A SPACE ILLUMINATED BY A LIGHT SOURCE.**

- Looper, S., Rodriguez-Puigvert, J., Siegwart, R., Cadena, C., Schmid, L. (2023, May). **3D vsg: Long-term semantic scene change prediction through 3D variable scene graphs.** In 2023 IEEE International Conference on Robotics and Automation (ICRA) (pp. 8179-8186). IEEE.

## 1.7 Manuscript organization

This thesis is organized as follows. In Chapter 2, we discuss about scalable Bayesian deep neural networks for supervised learning of single-view depth and we explore the importance of uncertainty quantification in robotics perception. In Chapter 3, we discuss about uncertainty of deep ensembles applied to different types of learning for single-view depth from colonoscopy images. In Chapter 4, we discuss the advantages of using a physical-model light-camera based to introduce the first single-view self-supervised method for depth learning. Finally, in Chapter 5, we summarize the conclusions of the thesis work.

# Chapter 2

# Bayesian deep neural networks for supervised learning of single-View depth

Understanding and quantifying uncertainty is crucial when using deep neural networks in real-world scenarios. The aim of this chapter is to explain how uncertainty can be quantified for deep neural networks and to evaluate the calibration of the captured uncertainty. We distinguish between two forms of uncertainty: aleatoric and epistemic. Aleatoric uncertainty arises from the inherent randomness or variability in the data, and represents the intrinsic noise of the system. Epistemic uncertainty, on the other hand, arises from either insufficient knowledge or a lack of data. However, as more data are collected and models refined, epistemic uncertainty can be reduced or even eliminated. To investigate these uncertainties, our approach focuses on modelling the aleatoric directly in the network and capturing the epistemic using two scalable techniques: MC dropout and deep ensembles. With respect to MC dropout, we explore the importance of dropout layers within the network architecture. Deep ensembles are a form of Bayesian model averaging, and in practice they can be used as an approximation of the full posterior.

## 2.1 Introduction

The quantification of the uncertainty is critical in robotics, in order to implement systems that are robust and reliable in real-world applications. Point estimators, which dominate the landscape of multi-view [Engel et al., 2017] and single-view [Fu et al., 2018, Godard et al., 2019] scene reconstruction, do not typically compute the full distribution but just the maximum-likelihood state. Higher-level decision blocks have no means to judge how accurate these estimates are, and hence its use for safe planning might be questionable. Uncertainty is often present in the formulation of model-based estimators (e.g., [Barfoot, 2017]), but much less in the training of deep learning models. Furthermore, learning-based approaches tend to overfit on standard datasets, which might lead us to assume a reasonable

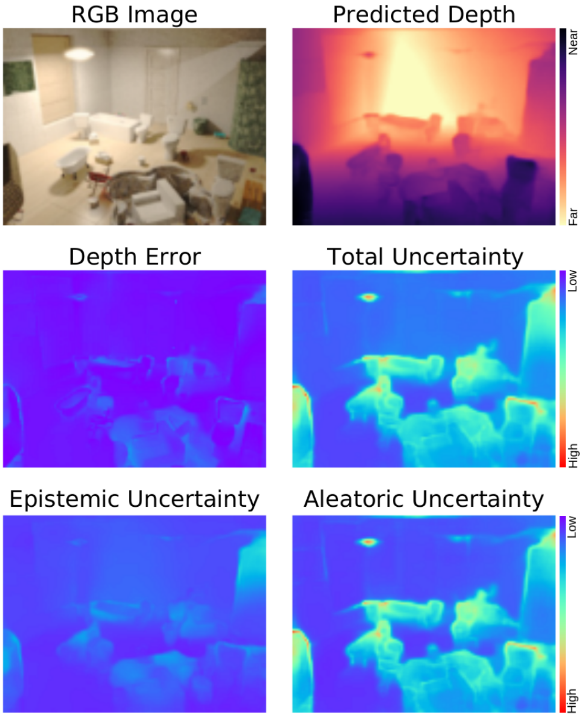


Figure 2.1: Bayesian single-view depth predicton for a SceneNet image. In the middle row the small depth error, and how the total uncertainty models it accurately. In the bottom row how epistemic and aleatoric sources are both significant and relevant for uncertainty quantification.

general performance while they are strongly biased. In such cases, their outputs should not be trusted in real-world applications, for which we would need generalizable and self-aware models. Bayesian learning is one of the approaches that can address such application challenges. In a supervised learning approach using depth annotations, we investigate epistemic and aleatoric uncertainties to determine the accuracy of the model's predictions and its suitability for use in robotic environments. In neural networks, uncertainty can stem from the input data or the network weights. For the latter, scalable approaches to Bayesian deep learning have shown to be effective to model uncertainty [Gustafsson et al., 2020]. Figure 2.1 shows the output of our Bayesian framework for supervised single-view depth learning. The small depth error is noticeable, but more importantly, the fact that this error is mostly coherent with the predicted uncertainty. Note also how the aleatoric uncertainty is greater in this case, but the epistemic uncertainty is still relevant for an accurate quantification. In this chapter, we provide a unified framework and a thorough evaluation of scalable uncertainty estimation methods, namely Monte Carlo (MC) dropout and deep ensembles, for supervised single-view depth learning with deep convolutional networks. We propose to apply MC dropout in the encoder, contrary to recent works [Gustafsson et al., 2020, Poggi et al., 2020] that apply it in the decoder. We demonstrate in our evaluations that, in the particular task of single-view depth supervised learning, the dropout in the encoder achieves a better performance than the dropout in the decoder. Such a result has relevant practical implications, as MC dropout has a lower memory footprint than deep ensembles. Finally, we also provide results in pseudo-RGBD ICP, a potential application for our single-view depth uncertainty models. In our experiments, we demonstrate that our uncertainty estimates are reasonably well calibrated and has significant potential to provide accurate and scaled motion estimates from monocular views.

## 2.2 Background and related work

### 2.2.1 Structure estimation and learning from images

Reconstructing a 3D scene from visual data has been addressed from a wide variety of perspectives, the more typical being based on multiple images either alone [Mur-Artal and Tardós, 2017, Engel et al., 2017] or fused with other sensors (e.g., visual-inertial setups [Qin et al., 2018]). However, multi-view and visual-inertial pipelines present two limitations: they require sufficiently textured scenes to find correspondences, and also sufficient motion for observability. Single-view depth estimation can help with these two issues, although it is significantly more challenging due to its ill-posed nature.

Following the seminal work of Saxena et al. [2005] on single-view depth, Eigen et al. [2014] were the first ones that used depth supervision to train a deep network for such task. Many works followed with different contributions: Laina et al. [2016] proposed deeper fully convolutional models. Fu et al. [2018] proposed a spacing-increasing depth discretization that learns depth from an ordinal regression perspective. Dijk and de Croon [2019] evaluated a self-trained networks to investigate which visual cues they use. From their conclusions, depth networks favor vertical positions and disregard obstacles in their apparent size. Similarly, we contribute to understanding the behavior of depth networks from a Bayesian perspective.

Several works have proposed self-supervised approaches, using photometric reprojection losses between stereo or multiple views [Godard et al., 2017, Zhou et al., 2017]. Self-supervised approaches still underperform compared to supervised ones. For this reason, we focus on supervised methods.

### 2.2.2 Bayesian deep learning

Bayesian deep learning combines the strengths of deep neural architectures with the uncertainty quantification of probabilistic (Bayesian) learning and inference methods. Regarding uncertainty, we must differentiate between what the model does not know and what is missing from the input data. Accordingly, the uncertainty sources can be classified into two: aleatoric and epistemic. Aleatoric (also referred to as statistical) uncertainty, refers to the variations caused by the realization of different experiments with stochastic components. In our models, it encodes the variability in the different inputs from the test data and hence cannot be reduced by increasing the amount of training data. Some models assume that the aleatoric uncertainty is homoscedastic; that is, it is independent of the input data. In this chapter, we train the network to predict the uncertainty for each input datum resulting in a heteroscedastic uncertainty model [Kendall and Gal, 2017, Der Kiureghian and Ditlevsen, 2009]. Epistemic (also known as systematic) uncertainty, represents the lack of knowledge of a trained model. This type of uncertainty is deeply related to the training data and the model ability to generalize. For example, epistemic uncertainty is high for out-of-distribution data or extrapolation in regions where training data was scarce. In Bayesian deep learning, the epistemic uncertainty can be estimated from the uncertainty in the model parameters, assuming that the model architecture is correct. In this case, epistemic uncertainty can be obtained using a prior distribution $p(\theta)$ over the neural network parameters $\theta$ and computing its posterior distribution $p(\theta|\mathcal{D})$ given a dataset $\mathcal{D}$ using Bayes rule: $p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}$. In general, this equation is intractable for state-of-the-art deep architectures, but there are several approaches to tackle this problem that we describe below.

**Variational inference (VI).**

VI proposes the use of a tractable approximation $q(\theta)$ to the posterior distribution $p(\theta|\mathcal{D})$. Mean-field variational inference assumes an isotropic Gaussian distribution for $q(\theta) \sim \mathcal{N}(\theta|\mu, \mathbf{I}\sigma)$. The parameters of the approximate distribution $q(\theta)$ are optimized by minimizing the KL-divergence between the approximate distribution and the true posterior $D(q||p)$. Mean-field variational inference with Gaussian approximation suffers from the soap-bubble effect, reducing the predictive performance as most samples fall in a ring. The Radial Bayesian Neural Networks [Farquhar et al., 2020] avoid that effect, but the distribution is biased towards the center resulting in uncertainty underestimation. Furthermore, VI methods are very sensitive to calibration and configuration. A natural-gradient VI method [Osawa et al., 2019] was introduced to improve the robustness of the optimization. However, it requires strong approximations of the Hessian, resulting in lower performance.

**Monte Carlo (MC) dropout.**

MC dropout can be used to approximate the posterior distribution, as proposed by Gal and Ghahramani [2016]. It can be considered as a specific case of VI, where the variational distribution includes a set of binary random variables that represent the corresponding unit to be turned off or dropped. The approximation makes the computation tractable and robust. MC dropout is able to approximate multimodal distributions. However, the epistemic distribution on the weight-space only has discrete support. Kendall and Gal [2017] presented a framework to combine both aleatoric and epistemic uncertainty, where MC dropout is used to obtain epistemic uncertainty, while the function mapping the aleatoric uncertainty is learned from the input data.

**Deep ensembles.**

Deep ensembles [Lakshminarayanan et al., 2017] involves training the same architecture many times optimizing some MAP loss, but starting from different random initialization of its parameters. Therefore, deep ensembles are not truly a Bayesian approach as the samples are distributed according to the different local optima. Conversely, these models in an ensemble perform reasonably well, even considering the small number of random samples considered in practice, as all of the models are optimized and have a high likelihood. Therefore, deep ensembles can be considered an approximate Bayesian model average, although, in practice, they can also be used as a rough posterior approximation. Contrary to MC dropout, where the model weights are shared between samples, in the case of deep ensembles, each *sample* is trained independently. Therefore the number of model parameters required grows linearly with the number of samples. Furthermore, deep ensembles also result in a distribution with discrete support on the weight-space.

### 2.2.3 Bayesian deep learning in computer vision

Evaluating uncertainty correctly is still in open discussion, as it is task-related. Mukhoti and Gal [2018] evaluated MC dropout for semantic segmentation and designed the metrics for such case. Similarly, Gustafsson et al. [2020] designed a framework to explore uncertainty metrics for semantic segmentation and depth completion, using MC dropout and deep ensembles. Ilg et al. [2018] compare different strategies and techniques for quantifying the uncertainty of optical flow. They also introduce a multi-hypothesis network based on a winner-takes-all loss function that penalizes the best hypothesis result. However, disentangling aleatoric and epistemic uncertainty could be an arduous task in multi-hypothesis approaches. The network that merge all hypotheses contains its own epistemic uncertainty that is not taking into account. Nevertheless, they show to be competitive in comparison to deep ensemble and MC dropout. For depth estimation, Yang et al. [2019] proposes a multinomial distribution to learn uncertainty based on discretizing the depth space.

For our case of single-view depth regression, Poggi et al. [2020] evaluated the uncertainty in self-supervised learning, which leverage photometric consistency between views. They observed that depth accuracy is improved by uncertainty estimation along the training paradigms. This thesis complements the ones mentioned in this section by evaluating uncertainty quantification in a supervised regression setting.

## 2.3 Bayesian single-view depth learning from supervised data

MC dropout and deep ensembles provide a sample representation of the posterior distribution over the network parameters. Here, we introduce a unified formulation to analyze the posterior and predictive distribution for these sample representations. We have particularized our framework for depth perception applications, although it could be extended to other tasks.

### 2.3.1 Architecture and loss

We adapt a U-Net [Ronneberger et al., 2015] encoder-decoder architecture as in [Godard et al., 2019, Poggi et al., 2020]. Our encoder is a Resnet18 [He et al., 2016] pre-trained in ImageNet [Russakovsky et al., 2015]. Table 2.1 summarizes our decoder architecture.

| Depth Decoder | | | |
|---|---|---|---|
| **layer** | **# filters** | **inputs** | **activation** |
| upconv5 | 256 | econv5 | ELU |
| iconv5 | 256 | ↑upconv5, econv4 | ELU |
| upconv4 | 128 | iconv5 | ELU |
| iconv4 | 128 | ↑upconv4, econv3 | ELU |
| depth_unc4 | 2 | iconv4 | - |
| upconv3 | 64 | iconv4 | ELU |
| iconv3 | 64 | ↑upconv4, econv3 | ELU |
| depth_unc3 | 2 | iconv3 | - |
| upconv2 | 32 | iconv3 | ELU |
| iconv2 | 32 | ↑upconv2, econv1 | ELU |
| depth_unc2 | 2 | iconv2 | - |
| upconv1 | 16 | iconv2 | ELU |
| iconv1 | 16 | ↑upconv1 | ELU |
| depth_unc1 | 2 | iconv1 | - |

Table 2.1: Decoder architecture. Kernels are always $3 \times 3$ with stride 1. ↑ stands for $2 \times 2$ nearest-neighbor upsampling.

Our training data $\mathcal{D} = \{\{I_1, d_1\}, \ldots, \{I_N, d_N\}\}$ is composed by $N$ supervised pairs, each pair $i \in \{1, \ldots, N\}$ containing a RGB image $I_i \in \{0, \ldots, 255\}^{w \times h \times 3}$ and its ground truth depth $d_i \in \mathbb{R}_{>0}^{w \times h}$. For a single input image, the network $f_\theta(I)$ outputs two channels: per-pixel depth $\widehat{d}(I)$ and uncertainty $\sigma_d(I)$. The later corresponds to aleatoric uncertainty, which can also be interpreted as heteroscedastic observation noise. We incorporate both output channels in a single loss per image by using a standard Laplace log-likelihood [Kendall and Gal, 2017]:

$$\mathcal{L}(\theta) = \frac{1}{w \cdot h} \sum_{j \in \Omega} \frac{||d[j] - \widehat{d}[j]||}{\sigma_d[j]} + \log \sigma_d[j] \tag{2.1}$$

where $j \in \Omega$ is the pixel index in the image domain $\Omega$.

For deep ensembles, the loss function is evaluated independently for each sample model as they are trained separately, resulting in $M$ sets of parameters $\{\theta_m\}_{m=1}^{M}$. Although the sample models are not drawn from the posterior distribution, it still can be considered an approximation in practice. Deep ensembles are especially suitable for our problem as we need to maintain the number of samples small to keep it tractable. Therefore, it is important that we are not wasting valuable resources in low probability models that might reduce the overall performance. In the case of MC dropout, the loss function can be used for approximate variational inference on the posterior distribution of the weights by training with dropout after every layer. The actual Monte Carlo phase is done by also performing random dropout at test time to sample from the variational distribution computed during training [Kendall and Gal, 2017]. This sampling at test time results again in a set of $\{\theta_m\}_{m=1}^{M}$ different parameters. This time they are all generated from the same trained model, resulting in a much lower memory and computational footprint compared to deep ensembles or other variational methods.

In practice, we found that adding dropout at every layer reduced the predictive performance considerably for our application, which is consistent with previous results [Mukhoti and Gal, 2018]. Therefore, in section 2.4.3, we study different configurations of dropout and compare their quality both in terms of depth error and uncertainty quantification.

### 2.3.2 Bayesian prediction of sample-based deep networks

The predictive distribution for a pixel depth can be computed by integrating over the model parameters. We use the same strategy for MC dropout and deep ensembles, as they both use sample representations of the model parameters:

$$\begin{aligned} p(d|I, \mathcal{D}) &= \int p(d|I, \theta) p(\theta|\mathcal{D}) d\theta \\ &\approx \sum_{m=0}^{M} p(d|I, \theta_m) \end{aligned} \tag{2.2}$$

As our architecture generates a Gaussian prediction for the pixel depth $\mathcal{N}(\widehat{d}, \sigma_d^2)$, the sample-based output is a mixture of Gaussians that can be approximated by a single Gaussian. In particular, for the $\{\theta_m\}_{m=1}^{M}$ model samples (MC dropout) or models (deep ensembles) with respective outputs $\widehat{d}(m)$ and $\sigma_d(m)$, we approximate the total predictive

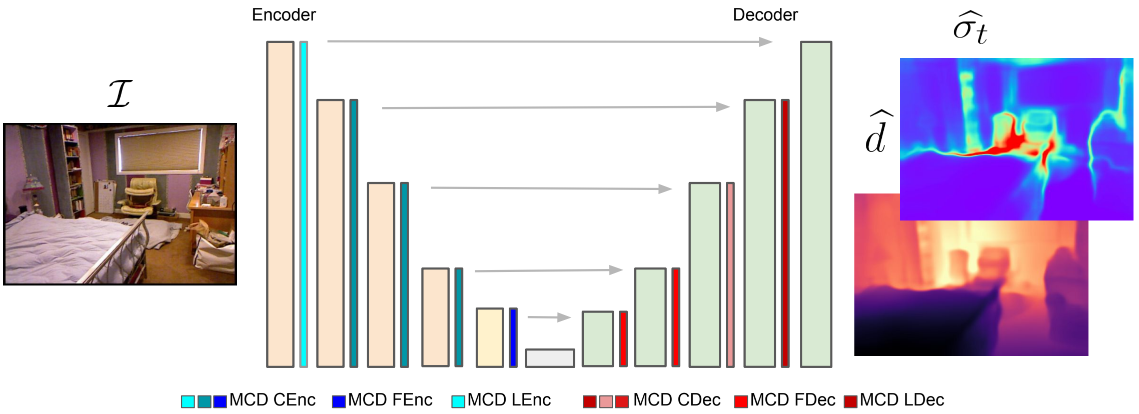


Figure 2.2: Variations of MC dropout in our experiments.

distribution per pixel as a Gaussian $p(d|I,\mathcal{D}) \approx \mathcal{N}(\widehat{d}_t, \sigma_t^2)$ with:

$$
\begin{aligned}
\widehat{d}_t &= \frac{1}{M}\sum_{m=1}^{M}\widehat{d}_m \\
\sigma_t^2 &= \underbrace{\frac{1}{M}\sum_{m=1}^{M}\left(\widehat{d}_t - \widehat{d}(m)\right)^2}_{\text{epistemic}} + \underbrace{\frac{1}{M}\sum_{m=1}^{M}\sigma_d^2(m)}_{\text{aleatoric}}
\end{aligned}
\tag{2.3}
$$

In the experiments section, we will show that identifying and quantifying the epistemic from the aleatoric uncertainty will be fundamental to finding the uncertainty source and improving the quality of the model and the predictions.

## 2.4 Experiments

### 2.4.1 Datasets

**SceneNet RGB-D dataset [McCormac et al., 2016]** contains photorealistic sequences of synthetic indoor scenes from general camera trajectories, along with their ground truth. Our models are trained over $210,000$ synthetic images of $700$ scenes and tested on $90,000$ images of $300$ different scenes. We chose this dataset as it provides a wide variety of viewpoints and scenes, challenging occlusions and different lighting conditions, which are relevant for the network generalization.

**NYU Depth V2 [Nathan Silberman and Fergus, 2012]** consists of $120,000$ RGB-D images in $464$ indoor scenes. For training, we use $36,253$ images of $249$ scenes as proposed by Lapdepth [Song et al., 2021]. We test our models in the official split of $654$ images [Eigen et al., 2014].

### 2.4.2 Metrics

We use the depth error metrics that are standard in literature: Absolute Relative difference: Normalize per pixel error according to the real depth, reducing the effect of large error with the distance Eq. 2.4, Square Relative difference: penalize larger square errors, Root mean square error (RMSE), Root mean square error Log ($RMSE_{Log}$) and $\delta < 1.25^i$ with $i \in \{1, 2, 3\}$ [Eigen et al., 2014]:

$$AbsRel = \frac{1}{w \cdot h} \sum_{j \in \Omega_i} \frac{|d[j] - \widehat{d}[j]|}{\widehat{d}[j]} \tag{2.4}$$

$$SqRel = \frac{1}{w \cdot h} \sum_{j \in \Omega_i} \frac{(d[j] - \widehat{d}[j])^2}{\widehat{d}[j]} \tag{2.5}$$

$$RMSE = \frac{1}{w \cdot h} \sum_{j \in \Omega_i} (d[j] - \widehat{d}[j])^2)^{1/2} \tag{2.6}$$

$$RMSE_{Log} = (\frac{1}{w \cdot h} \sum_{j \in \Omega_i} (\log d[j] - \log \widehat{d}[j])^2)^{1/2} \tag{2.7}$$

$$\delta < 1.25^i = \frac{1}{w \cdot h} \sum_{j \in \Omega_i} \max(\frac{d[j]}{\widehat{d}[j]}, \frac{\widehat{d}[j]}{d[j]}) < 1.25^i \tag{2.8}$$

For uncertainty, we use the Area Under the Calibration Error curve (AUCE) and Area Under the Sparsification Error curve (AUSE) [Gustafsson et al., 2020]. Since our methods output a Gaussian distribution $\mathcal{N}(\widehat{d}, \sigma^2)$ per pixel, we generate prediction intervals $\widehat{d} \pm \phi^{-1}(\frac{p+1}{2})\sigma$ of confidence level $p \in [0, 1]$ being $\phi$ the CDF of the standard normal distribution. In a perfectly calibrated model, the proportion of pixels for which the prediction intervals covers the ground truth coincides the confidence level. AUCE is an absolute uncertainty metric, we use AUSE, in terms of RMSE, as relative measure of uncertainty. This metric compares the ordering of the per-pixel uncertainties against the order of the per-pixel depth errors. The ordering should be similar for a well-calibrated uncertainty, as uncertain predictions will tend to have larger errors.

For the pseudo-RGBD Bayesian ICP in Section 2.4.4, we report the translational and rotational RMSE.

### 2.4.3 Bayesian single-view depth

We evaluate several variations of MC dropout and deep ensembles. Specifically, for MC dropout, we report depth and uncertainty metrics for dropouts at different layers and with $p = 0.3$ and $p = 0.5$, $p$ being the probability of an element to be zeroed.

**MC dropout.**

We consider seven variations (see Figure 2.2 for a summary plot): In the decoder, dropout after every convolutional layer, (**MC D**ropout **C**omplete **Dec**oder), the first two convolutional layers (**MC D**ropout **F**irst **Dec**oder), and the last convolution layer of the decoder (**MC D**ropout **L**ast **Dec**oder). And, in the encoder, dropout after every convolutional layer (**MC D**ropout **C**omplete **Enc**oder), the first convolutional layer (**MC D**ropout **F**irst **Enc**oder), and after the last convolution layer (**MC D**ropout **L**ast **Enc**oder). Finally, in NYU RGB-D v2 we also evaluate the effect of the dropout in all layers of the architecture (**MC D**ropout **C**omplete **Enc**oder-**C**omplete **Dec**oder).

**Deep ensembles.**

We examine uncertainty and depth by averaging an ensemble composed of a variable number of networks. We initialise the network weights with different seeds from a normal distribution $\mathcal{N}(0, 10^{-2})$.

**Results.**

Table 2.2 shows the depth and uncertainty metrics for the MC dropout variations and deep ensembles on SceneNet RGB-D. We observe that the uncertainty metric for deep ensembles outperforms all variants of MC dropout. This is due to the fact that deep ensembles are optimized to be close to a minimum. However, the depth error metrics are consistently better for MCD CEnc, which also show the second best uncertainty metrics.

The performance of the different MC dropout models varies significantly, which is a novel result of our analysis. We found that introducing dropout in all layers of the encoder (MCD CEnc) improves the results with respect to applying it to a few layers of the encoder (MCD FEnc, LEnc) or to the decoder (MCD FDec, LDec and CDec). Again, this result

| Model | M | Abs Rel | Sq Rel | RMSE | RMSE Log | $\delta < 1.25$ | $\delta < 1.25^2$ | $\delta < 1.25^3$ | AUSE |
|---|---|---|---|---|---|---|---|---|---|
| MCD CEnc p=0.3 | 64 | **0.1231** | **0.1222** | **0.6396** | **0.1982** | **0.8719** | **0.9635** | **0.9850** | 0.0985 |
| MCD CEnc p=0.5 | 64 | 0.1243 | 0.1228 | 0.6484 | 0.2023 | 0.8675 | 0.9615 | 0.9842 | 0.0968 |
| MCD FEnc p=0.3 | 64 | 0.1291 | 0.1326 | 0.6688 | 0.2047 | 0.8591 | 0.9591 | 0.9834 | 0.1210 |
| MCD FEnc p=0.5 | 64 | 0.1320 | 0.1363 | 0.6633 | 0.2062 | 0.8583 | 0.9586 | 0.9833 | 0.1229 |
| MCD LEnc p=0.5 | 64 | 0.1244 | 0.1245 | 0.6582 | 0.2010 | 0.8658 | 0.9607 | 0.9846 | 0.1219 |
| MCD LEnc p=0.3 | 64 | 0.1244 | 0.1248 | 0.6769 | 0.2000 | 0.8629 | 0.9595 | 0.9848 | 0.1327 |
| MCD CDec p=0.3 | 64 | 0.1316 | 0.1322 | 0.6781 | 0.2044 | 0.8567 | 0.9597 | 0.9841 | 0.1268 |
| MCD CDec p=0.5 | 64 | 0.1369 | 0.1378 | 0.6988 | 0.2080 | 0.8494 | 0.9579 | 0.9835 | 0.1323 |
| MCD FDec p=0.5 | 64 | 0.1263 | 0.1252 | 0.6690 | 0.2035 | 0.8604 | 0.9593 | 0.9841 | 0.1353 |
| MCD FDec p=0.3 | 64 | 0.1264 | 0.1263 | 0.6622 | 0.2016 | 0.8641 | 0.9604 | 0.9842 | 0.1304 |
| MCD LDec p=0.5 | 64 | 0.1336 | 0.1371 | 0.6866 | 0.2065 | 0.8568 | 0.9588 | 0.9833 | 0.1241 |
| MCD LDec p=0.3 | 64 | 0.1293 | 0.1303 | 0.6782 | 0.2042 | 0.8589 | 0.9593 | 0.9837 | 0.1285 |
| Deep ensembles | 18 | 0.1283 | 0.1244 | 0.6529 | 0.1993 | 0.8617 | 0.9613 | **0.9850** | **0.0838** |

Table 2.2: Depth and uncertainty metrics for several variations of MC Dropout and Deep ensembles in SceneNet RGB-Depth. Best results are boldfaced, second best ones are underlined.

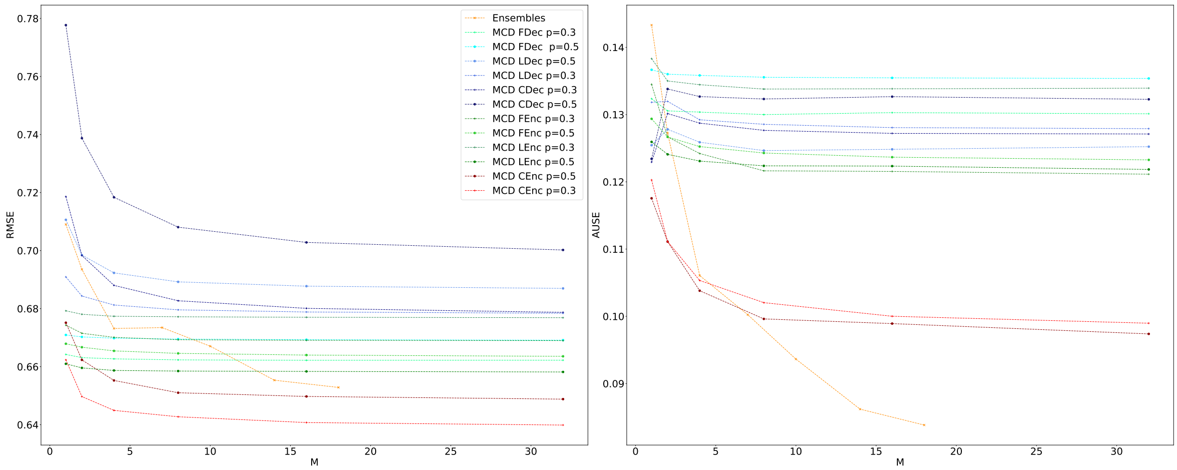


Figure 2.3: Comparison of MC dropout variations and deep ensembles for different numbers of forward passes $M$. Left: RMSE. Right: AUSE. The higher $M$ is, the better the performance, but with slight improvements for $M > 18$.

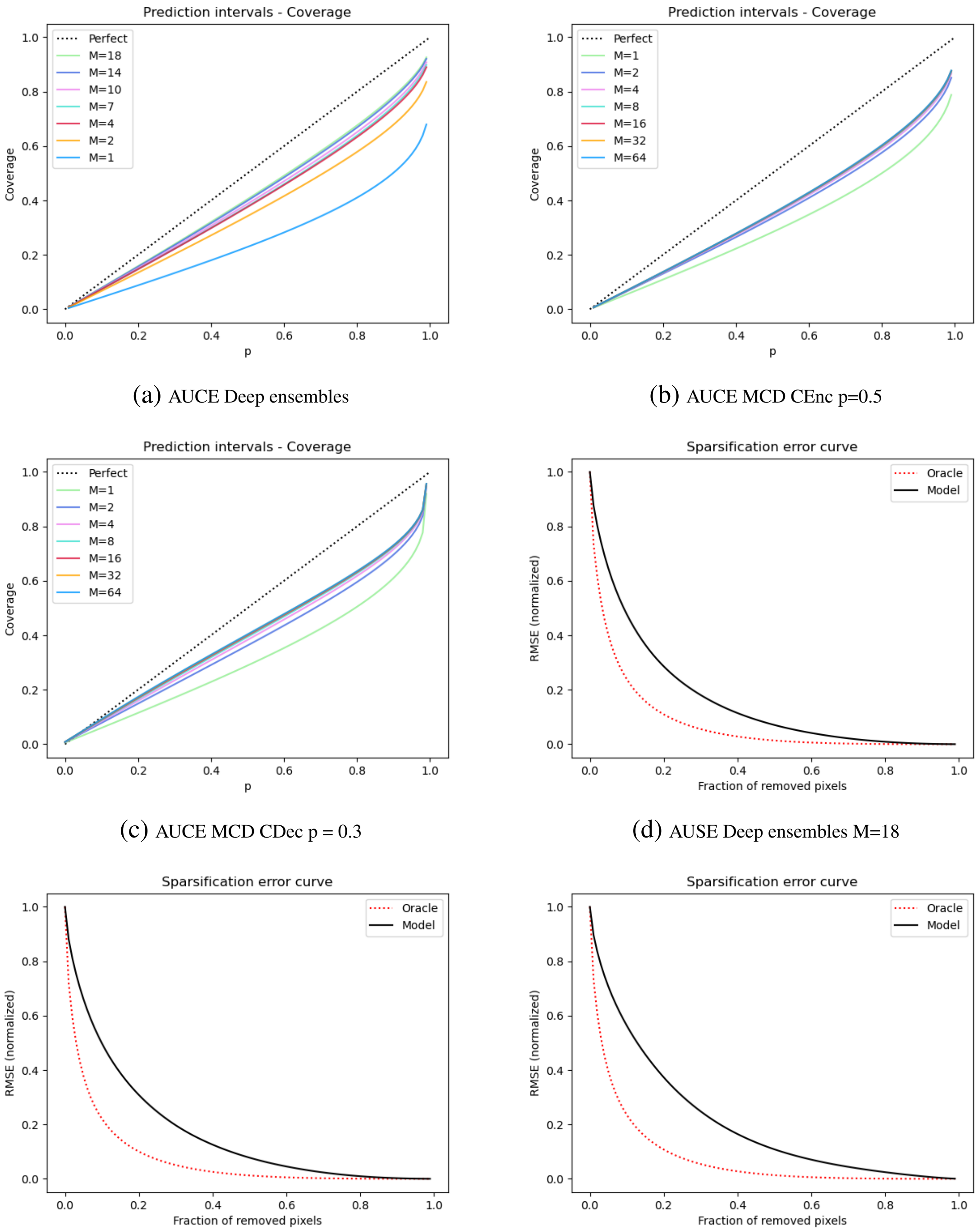
is of relevance as MC dropout is commonly applied in the decoder for depth estimation [Gustafsson et al., 2020, Poggi et al., 2020]. Our rationale for this result is as follows: we believe that applying MC dropout only in the decoder makes the network learn deterministic representations (the encoder is deterministic). In contrast, applying MC dropout in the image encoder allows us to learn probabilistic representations modeling uncertainty in the feature space. This seems to be a more appropriate choice for Bayesian image processing.

Our experiments show that there is small variations in the MC samples when we solely apply dropout close to the code, as done in MCD FDec or MCD Lenc. This results in worse calibrated models and also less satisfactory depth estimations. We also observed that applying MC dropout after the first and before the last layers of the network leads to poor performance (see MCD LDec, MCD FEnc).

Comparing the two dropout probabilities, $p = 0.3$ and $p = 0.5$, we observe that the depth prediction is usually better for $p = 0.3$ but the uncertainty is better calibrated for $p = 0.5$, as it introduces greater variability. One or the other should be preferred depending on the application.

MC dropout has a much smaller memory footprint than deep ensembles and our results indicate that the MCD CEnc performs similarly to ensembles, so it could be relevant option in certain applications. Specifically, only $56$ Mb are required for our MC dropout models, while the memory for deep ensembles grows linearly with the number of samples (for $M = 18$, around $1$ Gb). The run time grows linearly with the number of samples in both cases (around 3–4 ms per forward pass). All data stems from a NVIDIA GeForce RTX 3090 GPU.

Figure 2.3 shows the evolution of the metrics with the number of forward passes $M$. We can take $M = 1$ as the baseline, since there is no epistemic contribution. In general, predictions improve as $M$ increases. However, such improvements are hardly noticeable



(a) AUCE Deep ensembles

(b) AUCE MCD CEnc p=0.5

(c) AUCE MCD CDec p = 0.3

(d) AUSE Deep ensembles M=18

(e) AUSE MCD CEnc p=0.5 M=64

(f) AUSE MCD CDec p=0.3 M=64

Figure 2.4: Calibration curves (AUCE and AUSE) for MC dropout and deep ensembles.

for $M > 18$. Interestingly, for MCD CDec, the epistemic uncertainty results in a worse uncertainty calibration in comparison to a single forward pass.

Figure 2.4 shows the calibration error curves and sparsification error curves from where AUSE and AUCE were extracted. In these figures it can be seen that all models are overconfident, and the similarity between the sparsification error curves that leads to similar AUSE values.

Table 2.3 shows the metrics for depth and uncertainty on NYU Depth v2. We evaluate three MC dropout variations: MCD CEnc, MCD CDec, and MCD CEnc-CDec. In this dataset, deep ensembles show the best results in terms of uncertainty and depth metrics. As in SceneNet, applying dropout in the complete encoder (MCD CEnc) shows better results than applying it in the decoder (MCD CDec) or in the whole network (MCD CEnc-CDec). In this last case, the worse performance is caused by a too strong regularization effect. It is worth mentioning that the network is unable to recover the object boundaries and that the aleatoric uncertainty becomes diffuse for the high dropout rate case $p = 0.5$. This effect becomes more noticeable when the decoder contains dropout layers. Figure 2.8 shows results for one sample image. Observe that the areas with higher error correspond with the areas with higher uncertainty.

For all methods, increasing the number $M$ of forward passes improves the performance in uncertainty and depth (see Figure 2.7). We do not see a noticeable improvement, though, for $M > 18$ for deep ensembles and $M > 32$ for MC dropout.

Aleatoric uncertainty appears mainly in depth discontinuities, at object edges, and regions with sharp contrast in lighting (see Figure 2.5). As additional examples, Figure 2.6 shows highly uncertain depth predictions for three out-of-distribution images. It indicates high uncertainty values for unfamiliar objects not seen during training, like the dog and the door in the background of the first picture, the unrealistic patterns of the painting "Bedroom in Arles" by Van Gogh in the second one, and the person in the third one.

| Model | Abs Rel | Sq Rel | RMSE | RMSE Log | $\delta_1$ | $\delta_2$ | $\delta_3$ | AUCE | AUSE |
|---|---|---|---|---|---|---|---|---|---|
| Deep Ensembles | **0.1431** | **0.1052** | **0.5842** | **0.1973** | **0.8157** | **0.9596** | **0.9894** | 0.1302 | **0.1588** |
| MCD CENC-CDEC p=0.3 | 0.1560 | 0.1250 | 0.6332 | 0.2155 | 0.7867 | 0.9454 | 0.9849 | 0.1453 | 0.1703 |
| MCD CDec p=0.3 | 0.1574 | 0.1286 | 0.6307 | 0.2160 | 0.7850 | 0.9460 | 0.9851 | 0.1382 | 0.1770 |
| MCD CEnc p=0.3 | 0.1495 | 0.1173 | 0.6180 | 0.2090 | 0.8006 | 0.9491 | 0.9858 | 0.1436 | 0.1653 |
| MCD CENC-CDEC p=0.5 | 0.1642 | 0.1385 | 0.6592 | 0.2264 | 0.7671 | 0.9385 | 0.9824 | 0.1356 | 0.1668 |
| MCD CEnc p=0.5 | 0.1525 | 0.1220 | 0.6239 | 0.2127 | 0.7942 | 0.9467 | 0.9848 | 0.1377 | 0.1720 |
| MCD CDec p=0.5 | 0.1578 | 0.1291 | 0.6326 | 0.2164 | 0.7861 | 0.9457 | 0.9843 | **0.1286** | 0.1716 |

Table 2.3: Depth and uncertainty metrics for several variations of MC Dropout and Deep ensembles in NYU Depth V2. Best results are boldfaced, second best ones are underlined.

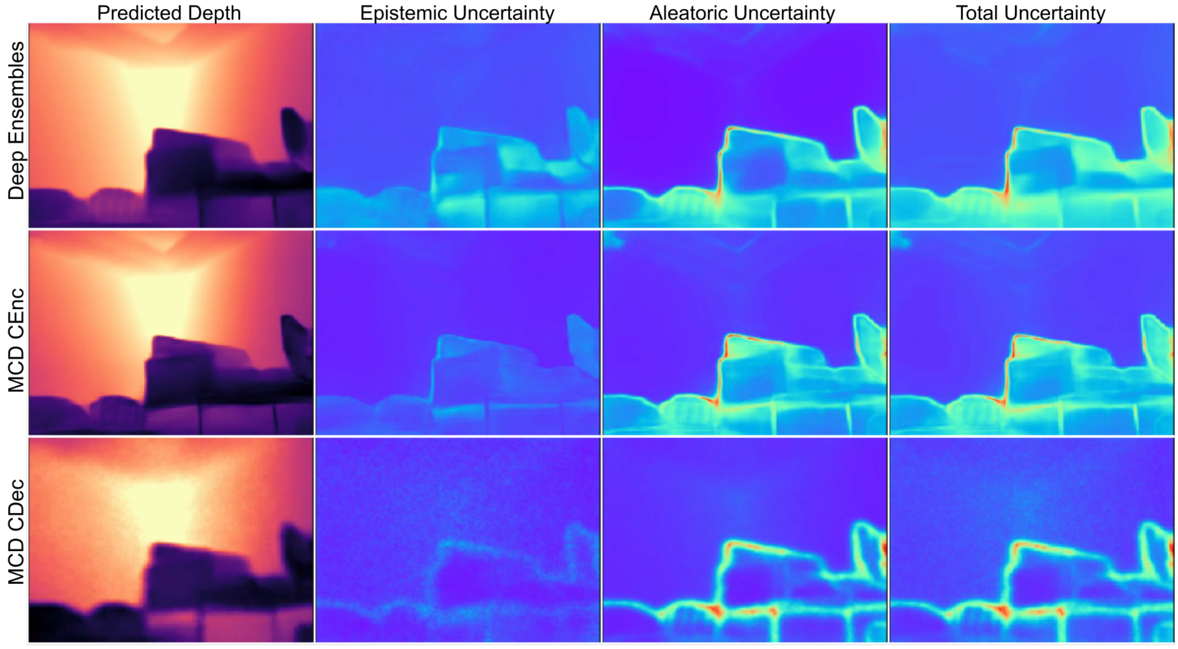


Figure 2.5: Depth and uncertainty predictions in a SceneNet RBG-D image.

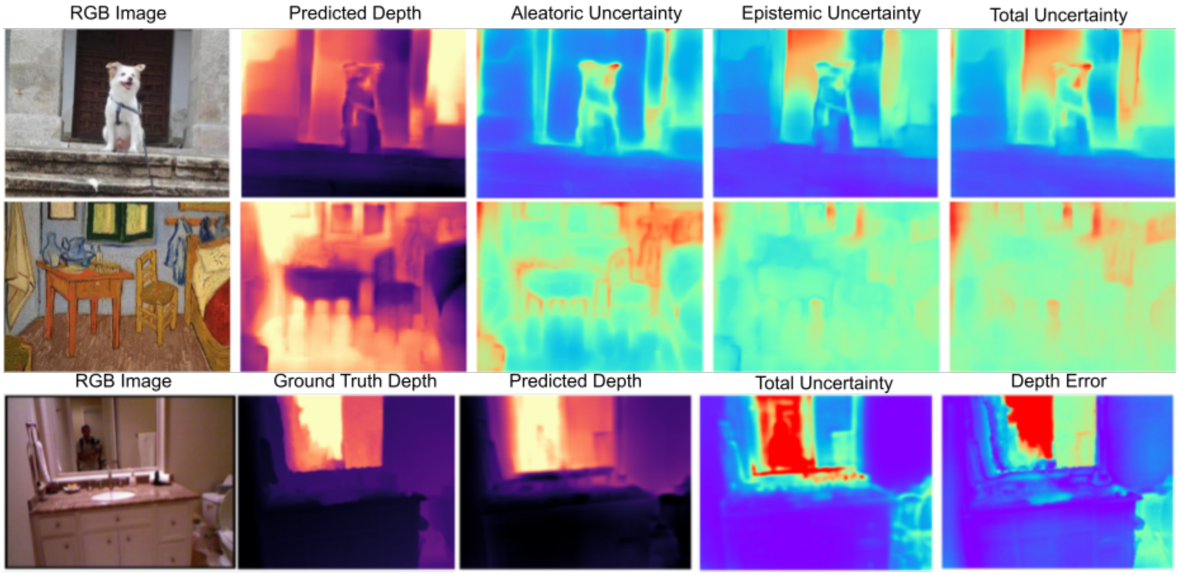


Figure 2.6: Predictions for three out-of-distribution images, showing high uncertainty for unfamiliar objects and textures. Top row: the aleatoric uncertainty is large in depth discontinuities and object boundaries, and the epistemic one concentrates in unknown patterns (the dog and the door). Middle row: aleatoric and epistemic uncertainties are both large due to large differences in appearance with respect to the training data. Bottom row: prediction in NYU Depth V2 image by an ensemble trained in SceneNet. The uncertainty is only very high for the human, not present in the training data.

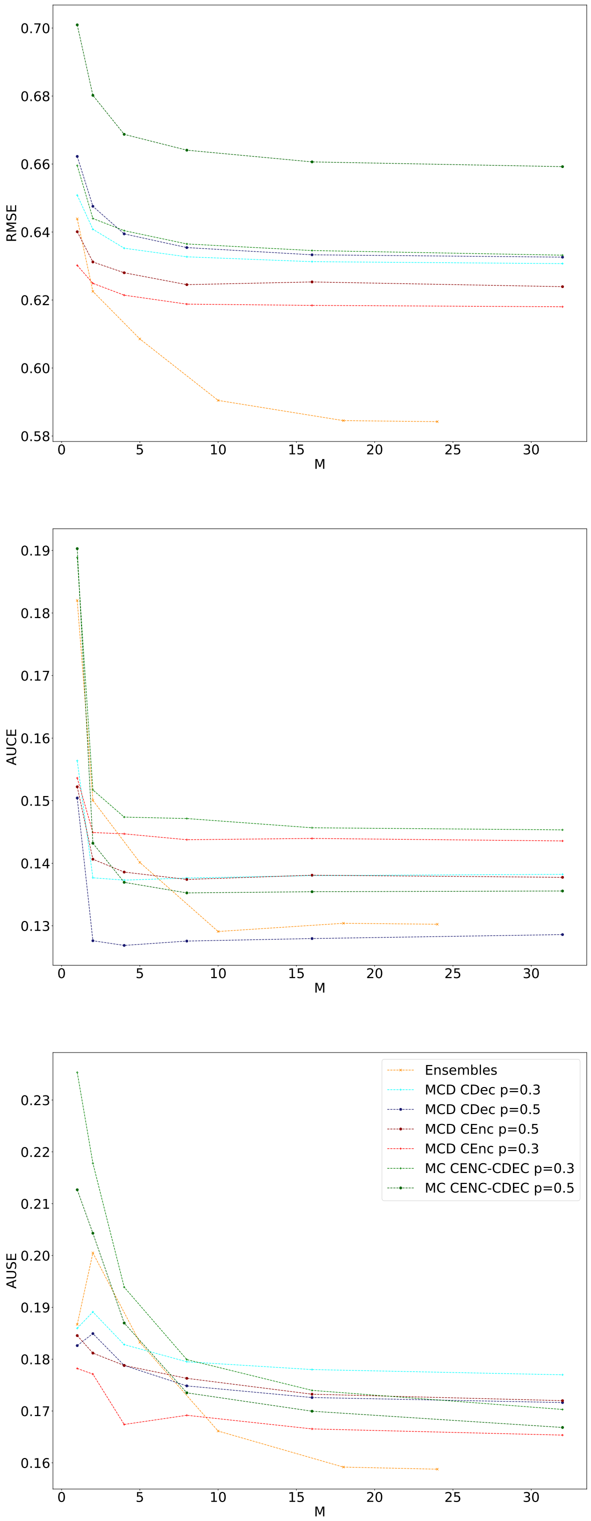


Figure 2.7: Comparison of MCD CDec, CEnc and Deep ensembles for forward passes $M$. Up: RMSE. Center: AUSE. Down: AUCE. The higher $M$ is, the better the performance, but with slight improvements for $M > 18$.

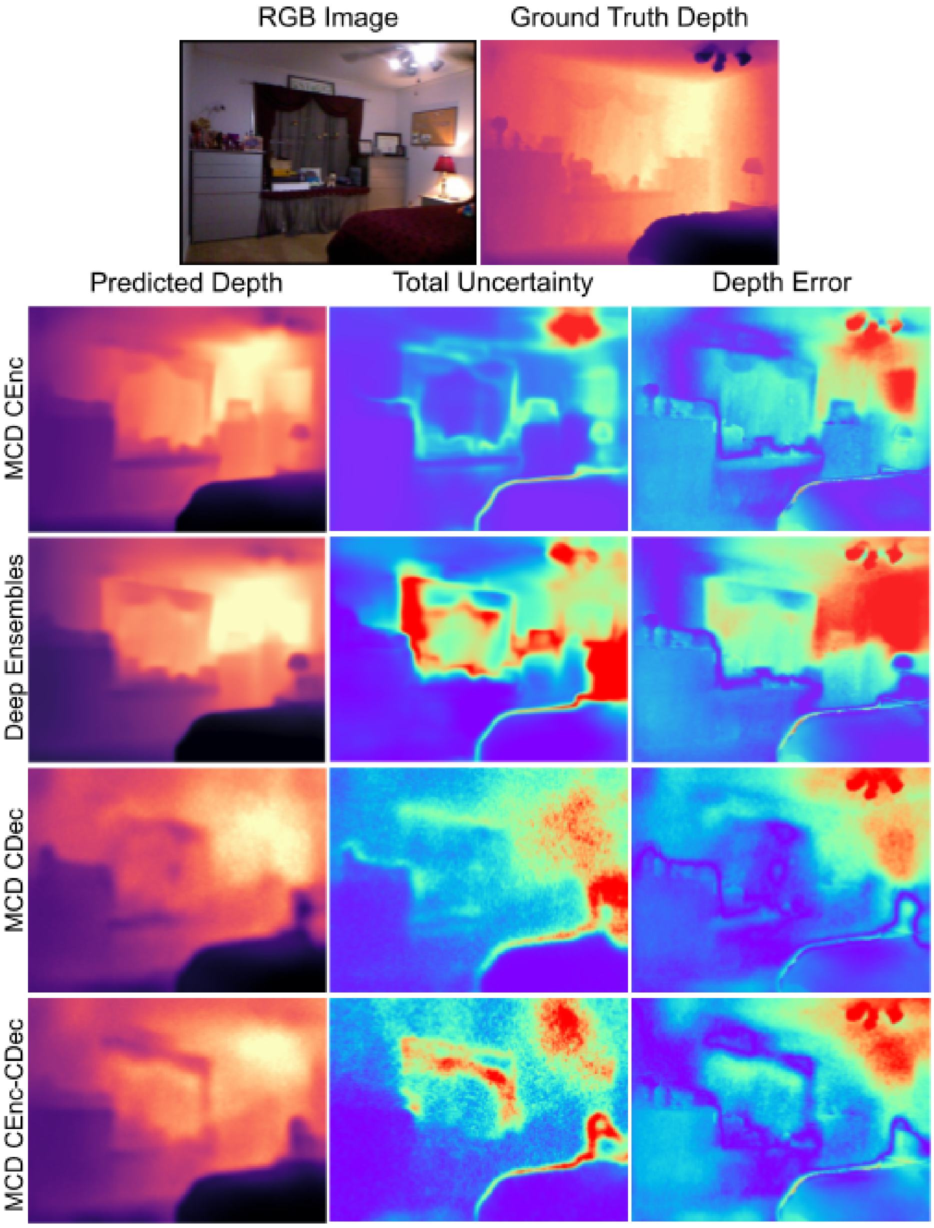


Figure 2.8: Depth and uncertainty results in NYU Depth v2 for MC dropout variations and deep ensembles. Top Row: input image and ground truth depth. First column: predicted depth. Second column: predicted uncertainty. Third column: depth error. Colors are equalized per image for better visualization.

### 2.4.4 Bayesian pseudo-RGBD ICP

In this section, we evaluate the application of Bayesian depth neural networks for two-view relative motion. Relative motion can be directly computed from deep neural networks where the inductive bias can be used to estimate absolute values even for uncalibrated cameras and blurry or poorly illuminated images. But geometric methods are more precise in a multi-view setting with a known baseline for scale disambiguation and sensible features can be tracked between multiple views [Zhou et al., 2020b]. In this work, we present a mixed approach where we use a two-view geometric method (ICP) to accurately compute the relative motion based on the depth prediction from neural networks. Thus, our approach is able to work with poor quality images and provide an unambiguous 3D transformation.

Our proposal leverages the depth predicted by a network to augment monocular images into what we call pseudo-RGBD views, and then aligns them using Iterative Closest Point (ICP). Similar ideas were proposed recently by Tiwari et al. [2020] and Luo et al. [2020]. Differently from us, they rely on Structure from Motion [Schonberger and Frahm, 2016] or visual SLAM [Mur-Artal and Tardós, 2017] to estimate the motion from the pseudo-RGBD views. Also, we use depth uncertainty for a more informed point cloud alignment, specifically excluding highly uncertain points from ICP.

| **Percentile** | .30 | .50 | .75 | .90 | .95 | .99 | 1.00 |
|---|---|---|---|---|---|---|---|
| **RMSE t** [m] | 0.238 | 0.216 | 0.188 | 0.182 | **0.179** | 0.190 | 0.190 |
| **RMSE r** [°] | 1.992 | 1.911 | 1.936 | **1.847** | 1.888 | 1.912 | 1.891 |

Table 2.4: ICP errors for percentiles .30, .50, .75, .90, .95, .99, 1.00. Best are boldfaced, second best underlined.

Our experimental setup is as follows. We selected 1408 random image pairs, separated by at least 4 frames, from SceneNet RGB-D. We excluded pairs with large areas without ground-truth depth (e.g., windows) and small overlap (rotations larger than 60°). We kept the image pairs for which there is sufficient evidence that ICP converged.

For each pair, we back-projected the estimated depth distributions into a point clouds and applied ICP to the following percentiles of the most certain points according to our estimation: .30, .50, .75, .90, .95, .99, and 1.00 (the percentile 1.00 corresponds to the full point clouds). We used our deep ensemble model, as it showed the best uncertainty calibration in Table 2.2.

Figure 2.9 illustrates our hypothesis with an example. The left point cloud is the original one, and the right one corresponds to the percentile .90. The points highlighted in red represent the 10% most uncertain points according to our uncertainty estimation, and clearly correspond to highly erroneous ones, as they lie on depth discontinuities. Removing such points from ICP will improve its accuracy.

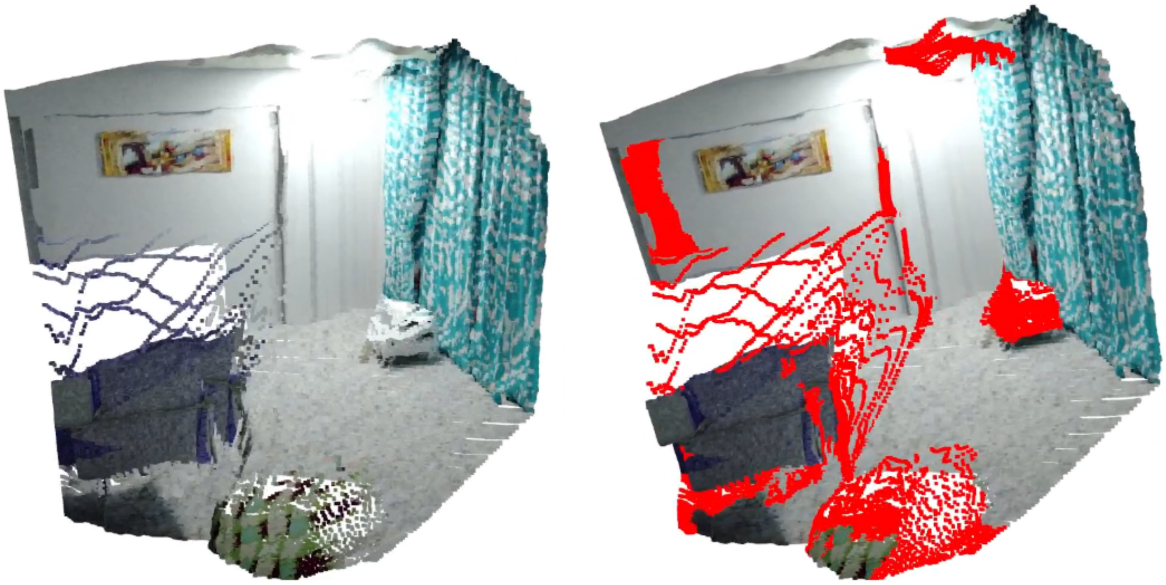


Figure 2.9: Left: Single-view reconstruction. Right: Same reconstruction, with the $10\%$ most uncertain points plotted in red. These points corresponds to spurious or high error points that will degrade the performance of ICP.

Table 2.4 shows the translational and rotational errors for all evaluated pairs. As motivated before, removing the most uncertain points (corresponding in well calibrated models to the most errouneous ones) reduces the estimation errors. The best results are for percentiles $.90$ and $.95$ ($1.00$ corresponds to the original point cloud). When the percentage of points removed is higher the error grows. This effect becomes obvious for percentiles $.30$ and $.50$, for which $70\%$ and $50\%$ of the points with the highest uncertainty were removed respectively. In these cases, the number of points removed is too large and the estimation becomes less accurate.

## 2.5 Conclusions

In this chapter, we evaluated MC dropout and deep ensembles as scalable Bayesian approaches to uncertainty quantification for single-view supervised depth learning.

We demonstrate empirically that using MC dropout in the encoder outperforms other variations used in the literature, which is a result of practical relevance. The placement of dropout in the architecture indeed has a significant effect in the estimation of depth and uncertainty. As a second conclusion of our analysis, deep ensembles have the best calibrated uncertainty estimations. However, applying dropout in the encoder performs only slightly worse than deep ensembles. As MC dropout needs much less memory than deep ensembles, it may be considered the Bayesian approach with more potential for applications. In our experimental results, we also show the application of Bayesian depth networks to pseudo-RGBD ICP, with the result that relative transformation can be improved by excluding the points with highest uncertainties.

# Chapter 3

# On the uncertain single-view depths in colonoscopies

Colonoscopy images, due to their complex nature and the environment they capture, are notoriously difficult for computer vision to interpret and analyze. The presence of folds or haustra results in discontinuities. Furthermore, reflections from the wet surfaces inside the colon create specular light, adding another dimension of complexity to understanding the images. This chapter explores scalable Bayesian deep networks to predict depth and uncertainty in these challenging images under different learning paradigms. Specifically, we first benchmark thoroughly, in synthetic data, supervised and self-supervised learning approaches in the colonoscopic domain. We address the problem of domain change and demonstrate that deep ensembles models trained on synthetic data can be transferred adequately to similar real domains. We propose a novel uncertainty aware teacher-student method that models the uncertainty of the teacher in the learning pipeline. Furthermore, we quantify the generalization of depth and uncertainty to real scenarios.

## 3.1 Introduction

Depth perception inside the human body is one of the cornerstones to enable automated assistance tools in medical procedures (e.g. virtual augmentations and annotations, accurate measurements or 3D registration of tools and interest regions) and, in the long run, the full automation of certain procedures and medical robotics. Monocular cameras stand out as very convenient sensors, as they are minimally invasive for in-vivo patients, but estimating depth from colonoscopy images is a challenge. Multi-view approaches are accurate and robust in many applications outside the body, e.g. Schonberger and Frahm [2016], but assume certain rigidity, texture and illumination conditions that are not fulfilled in in-body images. As mentioned in Chapter 2, single-view 3D geometry is ill-posed, since infinite 3D scenes can explain a single 2D view [Hartley and Zisserman, 2003]. As discussed in such chapter, deep

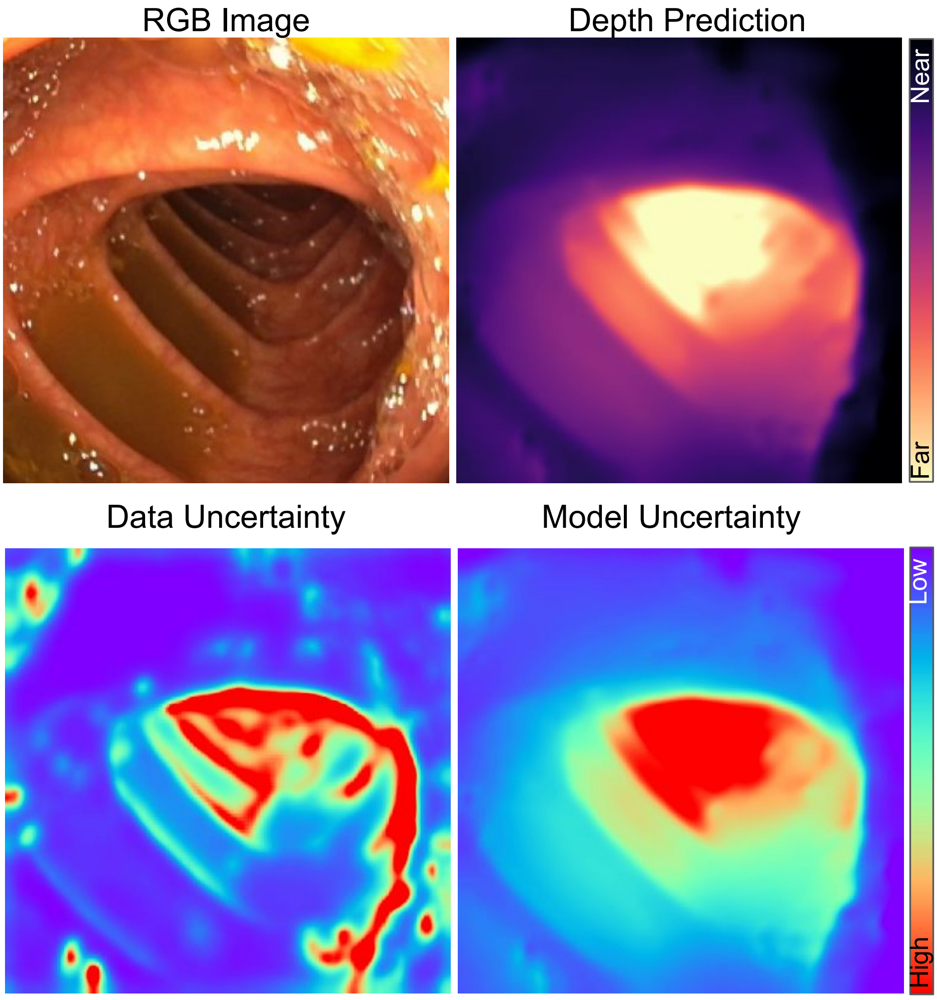


Figure 3.1: Depth and uncertainty predictions for a colonoscopy image. Dark/bright colors stands for near/far depths, blue/red stands for low/high uncertainties. Note the higher uncertainties in darker and farther areas and in reflections.

neural networks have shown impressive results in last years [Eigen et al., 2014, Fu et al., 2018, Godard et al., 2019]. However, the vast majority of deep learning models lack any metric or intuition about their predictive accuracy. In a critical environment such as the inside of the human body, uncertainty quantification is essential. Specifically, in medical robotics it allows us to properly account for uncertainty in control and decision making, and in SLAM [Durrant-Whyte and Bailey, 2006] to safely navigate inside the body. It also provides confidence intervals in in-body measurements (e.g., polyps), which is valuable for doctors to decide how to act. Uncertainty quantification is, a must-have for robust, interpretable and safe AI systems. Bayesian deep learning perfectly combines the fields of deep learning and uncertainty quantification in a sound and grounded approach. However, for high-dimensional deep networks, accurate Bayesian inference is intractable. Only bootstrapping methods such as deep ensembles [Lakshminarayanan et al., 2017] have shown to produce well-calibrated uncertainties in many computer vision tasks at reasonable cost [Gustafsson et al., 2020].

## 3.2 Preliminaries and related work

**Bayesian deep learning** is a form of deep learning that performs probabilistic inference on deep network models. This enables uncertainty quantification for the model and the predictions. For high-dimensional deep networks, Bayesian inference is intractable and some approximate inference methods such relying on variational inference or Laplace approximations might perform poorly. In practice, sampling methods based on bootstrapping, such as deep ensembles [Lakshminarayanan et al., 2017], or Monte Carlo, like MC dropout [Gal and Ghahramani, 2016], have shown to be the most scalable, reliable and efficient approaches for depth estimation and other computer vision tasks [Gustafsson et al., 2020]. In particular, deep ensembles have shown to perform extremely well even with a reduced number of samples, because each random sample of the network weight is optimized using a maximum a posteriori (MAP) loss $\mathcal{L}_{MAP} = \mathcal{L}_{LL} + \mathcal{L}_{prior}$, resulting in a high probability sample. The MAP loss requires a prior distribution, which unless otherwise stated, we assume to be a Gaussian distribution over the weights $\mathcal{L}_{prior} = ||\theta||^2$. Similarly as chapter 2, for the data likelihood $\mathcal{L}_{LL}$, we use a loss function based on the Laplace distribution for which the predicted mean $\mu(x)$ and the predicted scale $\sigma(x)$ come from the network described in Section 3.3, with two output channels [Kendall and Gal, 2017]. The variance associated with the scale term represents the uncertainty associated with the data, also called *aleatoric uncertainty* or $\sigma_a(x)$. Furthermore, in deep ensembles, the variance in the prediction from the multiple models of the ensemble is the uncertainty that is due to the lack of knowledge in the model, which is also called *epistemic uncertainty* or $\sigma_e(x)$. For example, data uncertainty might appear in poorly illuminated areas or with lack of texture, while model uncertainty arises from data that is different from the training dataset. Note that while the model uncertainty can be reduced with larger training datasets, data uncertainty is irreducible. Model uncertainty is particulary relevant to address domain changes. To illustrate that, in Section 3.5 we present results of models trained on synthetic data and tested on real data.

**Single-View Depth Learning** has demonstrated a remarkable performance recently. Some methods rely on accurate ground truth labels at training [Eigen et al., 2014, Fu et al., 2018, Song et al., 2021], which is not trivial in many application domains. Self-supervision without depth labels was achieved by enforcing multi-view photometric consistency during training [Zhou et al., 2017, Zhan et al., 2018, Godard et al., 2019]. In the medical domain, supervised depth learning was addressed by Visentini-Scarzanella et al. [2017] with autoencoders and by [Shen et al., 2019] with GANs, both using ground truth from phantom models. Other works based on GANs were trained with synthetic models [Chen et al., 2019b, Mahmood et al., 2018, Mahmood and Durr, 2018, Rau et al., 2019], and  [Cheng et al., 2021] added a temporal consistency loss.

Self-supervised learning is a natural choice for endoscopies to overcome the lack of depth labels on the target domain [Sharan et al., 2020, Ozyoruk et al., 2021, Recasens et al., 2021]. Although depth or stereo are not common for in-vivo procedures, several works use them for training [Luo et al., 2019, Xu et al., 2019, Huang et al., 2021]. Others train in phantoms [Turan et al., 2018] or synthetic data [Freedman et al., 2020, Hwang et al., 2021], facing the risk of not generalizing to the target domain. We study the limits of such generalization. SfM supervision was addressed by Liu et al. [2020] using siamese networks and by Widya et al. [2021] using GANs. Note that none of these references address uncertainty quantification, which we cover in this work. Kendall and Gal [2017] combine epistemic and aleatoric uncertainty by using a MC Dropout approximation of the posterior distribution. This approach obtains pixel-wise depth and uncertainty predictions in a supervised setting. Traditional self-supervised losses to regress depth are limited due to the aleatoric uncertainty of input images [Li et al., 2021]. Poggi et al. [2020] address such problem by introducing a teacher-student architecture to learn depth and uncertainty. As key advantages, teacher-student architectures provide aleatoric uncertainty for depth and avoid photometric losses and pose regression networks, which are frequently unstable. In this chapter, we present the evaluation for supervised and self-supervised approaches in colonoscopies images and propose a novel teacher-student approach that includes teacher uncertainty during training. Among the scalable Bayesian methods for single-view depth prediction, deep ensembles show the best calibrated uncertainty as demonstrated at Chapter 2 Rodríguez-Puigvert et al. [2022] and hence we choose them as our model.

## 3.3 Supervised learning using deep ensembles

Let our dataset $\mathcal{D} = \{\{\mathcal{I}_1, d_1\}, \ldots, \{\mathcal{I}_N, d_N\}\}$ be composed by $N$ samples, where each sample $i \in \{1, \ldots, N\}$ contains the input image $\mathcal{I}_i \in \{0, \ldots, 255\}^{w \times h \times 3}$ and per-pixel

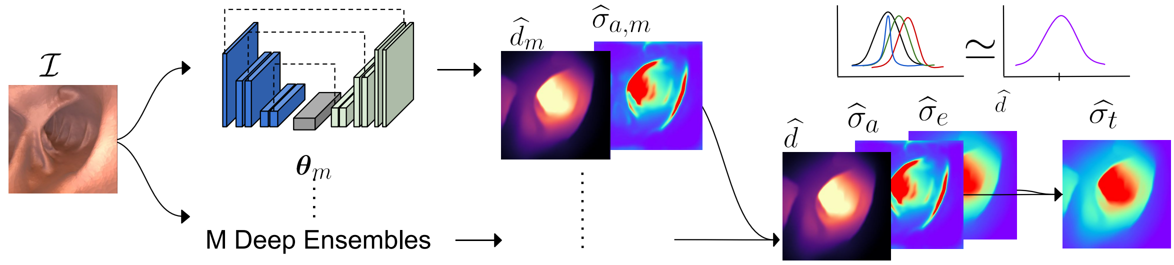


Figure 3.2: Forward propagation of supervised deep ensembles. Our deep ensembles model a Gaussian distribution $N(\widehat{d}, \widehat{\sigma}_t^2)$. $\widehat{d}$ comes from averaging all ensembles depth output and $\widehat{\sigma}_t^2$ from joining data and model uncertainties

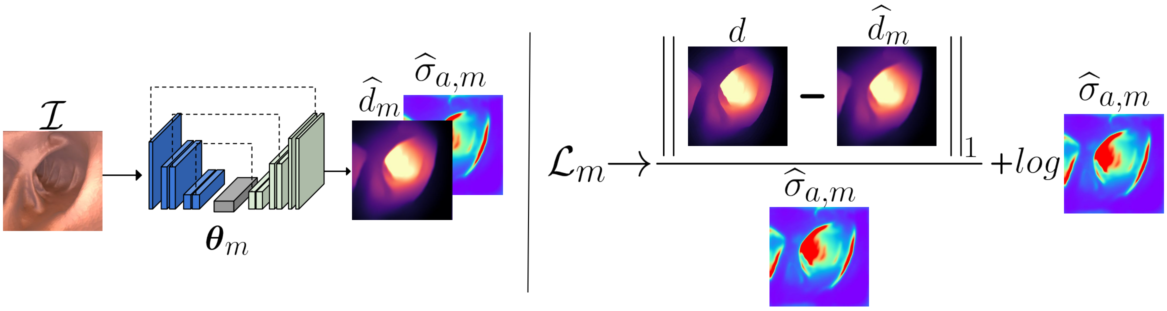


Figure 3.3: Supervised training of a single ensemble $m$. Depth $\widehat{d}_m$ and aleatoric uncertainty $\sigma_{a,m}$, estimation for a target image $\mathcal{I}$. We define graphically the loss for a single ensemble. $\widehat{d}_m$ depth prediction, $d$ depth target and $\sigma_{a,m}$ the aleatoric uncertainty of the ensemble.

depth labels $d_i \in \mathbb{R}^{w\times h}_{>0}$. Regarding our network, we use an encoder-decoder architecture with skip connections, inspired by Monodepth2 [Godard et al., 2019], with two output layers. Thus, for every new image $\mathcal{I}$ the network predits its pixel-wise depth $\widehat{d}(\mathcal{I},\boldsymbol{\theta}) \in \mathbb{R}^{w\times h}_{>0}$ and data variance $\widehat{\sigma}^2_a(\mathcal{I},\boldsymbol{\theta}) \in \mathbb{R}^{w\times h}_{>0}$. As commented in Section 3.2, we use a MAP loss, with $\mathcal{L}_{prior} = ||\boldsymbol{\theta}||^2$ and :

$$\mathcal{L}_{LL} = \frac{1}{w \cdot h} \sum_{\boldsymbol{j}\in\Omega_i} \left( \frac{||d[\boldsymbol{j}] - \widehat{d}[\boldsymbol{j}]\,||_1}{\widehat{\sigma}_a[\boldsymbol{j}]} + \log \widehat{\sigma}_a[\boldsymbol{j}] \right) \tag{3.1}$$

where $[\cdot]$ is the sampling operator and $\boldsymbol{j} \in \Omega$ refers to the pixel coordinates in the image domain $\Omega$. The per-pixel depth labels $d$ can be obtained from ground truth depth $d^{GT}$ or from SfM 3D reconstructions $d^{SfM}$ [Schonberger and Frahm, 2016]. Figure 3.3 presents the training of a single ensemble. A *deep ensemble model* is composed by $M$ networks with weights $\{\theta_m\}^M_{m=1}$, each of them trained separately starting from different random seeds. We denote as $(\widehat{d}_m, \widehat{\sigma}^2_{a,m})$ the output of the $m^{th}$ ensemble (see Figure 3.2). We obtain the mean depth of the ensemble $\widehat{d}$ and its epistemic uncertainty $\widehat{\sigma}^2_e$ using the total mean and variance of the full model. The total uncertainty $\widehat{\sigma}^2_t = \widehat{\sigma}^2_a + \widehat{\sigma}^2_e$ combines the data $\widehat{\sigma}^2_a$ and model $\widehat{\sigma}^2_e$ uncertainties which results from the law of total variance.

$$\widehat{d} = \frac{1}{M}\sum_{m=0}^{M} \widehat{d}_m, \;\; \widehat{\sigma}^2_a = \frac{1}{M}\sum_{m=0}^{M} \widehat{\sigma}^2_{a,m}, \;\; \widehat{\sigma}^2_e = \frac{1}{M}\sum_{m=0}^{M} \left(\widehat{d} - \widehat{d}_m\right)^2 \tag{3.2}$$

## 3.4 Self-supervised learning using deep ensembles

Self-supervised methods aim at learning *without* depth labels, the training data being $\mathcal{D} = \{\mathcal{I}_1, \ldots, \mathcal{I}_N\}$ and the supervision coming from multi-view consistency. For each instance $m$ of a deep ensemble, two deep networks are used [Godard et al., 2019].

The first one learning depth and a photometric uncertainty parameter $\widehat{u}$ and the second one learning to predict relative camera motion. We use a pseudo-likelihood for the loss

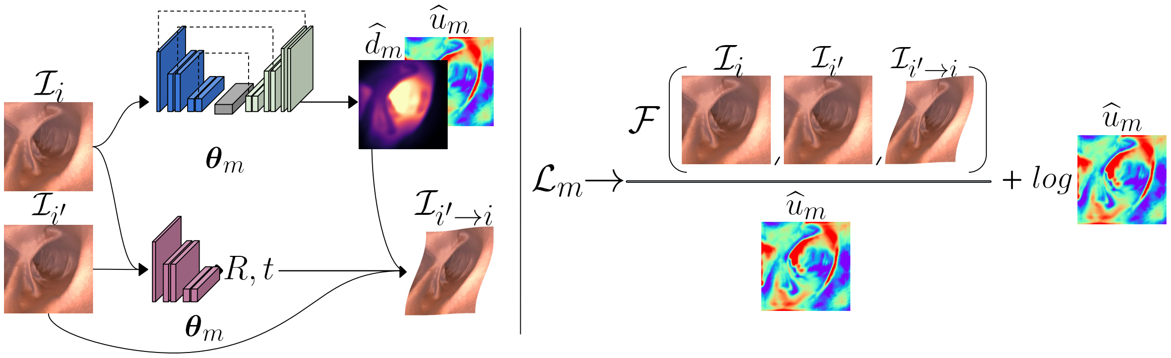


Figure 3.4: Self-supervised training of deep ensembles.

function, that uses both networks for photometric consistency:

$$\mathcal{L}_{LL,m} = \frac{1}{w \cdot h} \sum_{\boldsymbol{j} \in \Omega_i} \left( \frac{\mathcal{F}_p[\boldsymbol{j}]}{\widehat{u}_m[\boldsymbol{j}]} + \log \widehat{u}_m[\boldsymbol{j}] \right) \tag{3.3}$$

where $\mathcal{F}_p$ is the photometric residual and $\widehat{u}_m$ an uncertainty prediction (See Figure 3.4) The photometric residual $\mathcal{F}_p[\boldsymbol{j}]$ of pixel $\boldsymbol{j}$ in a target image $\mathcal{I}_i$ is the minimum –between the warped images $\mathcal{I}_{i'\to i}$ from the previous and posterior images $\mathcal{I}_{i'}$ to the target one $\mathcal{I}_i$– of the sum of the photometric reprojection error and Structural Similarity Index Measure (SSIM) [Wang et al., 2004]:

$$\mathcal{F}_p[\boldsymbol{j}] = \min((1-\alpha)\|\mathcal{I}_i[\boldsymbol{j}] - \mathcal{I}_{i'\to i}[\boldsymbol{j}]\|_1 + \frac{\alpha}{2}(1 - \mathrm{SSIM}(\mathcal{I}_i, \mathcal{I}_{i'\to i}, \boldsymbol{j})) \tag{3.4}$$

being $\alpha \in [0,1]$ the relative weight of the addends; and $\mathcal{I}_i[\boldsymbol{j}]$ and $\mathcal{I}_{i'\to i}[\boldsymbol{j}] = \mathcal{I}_{i'}[\boldsymbol{j}']$ the color values of pixel $\boldsymbol{j}$ of the target image $\mathcal{I}_i$ and of the warped image $\mathcal{I}_{i'\to i}$. To obtain this latter term, we warp every pixel $\boldsymbol{j}$ from the target image domain $\Omega_i$ to that of the source image $\Omega_{i'}$ using:

$$\boldsymbol{j}' = \pi\Big(\mathbf{R}_{i'i}\pi^{-1}(\boldsymbol{j}, \widehat{d}_i[\boldsymbol{j}]) + \mathbf{t}_{i'i}\Big) \tag{3.5}$$

$\mathbf{R}_{i'i} \in \mathrm{SO}(3)$ and $\mathbf{t}_{i'i} \in \mathbb{R}^3$ are the rotation and translation from $\Omega_i$ to $\Omega_{i'}$, and $\pi$ and $\pi^{-1}$ the projection and back-projection functions (3D point to pixel and vice versa). In this case, the prior loss also incorporates an edge-aware smoothness term $\mathcal{F}_s$, regularizing the predictions [Godard et al., 2019]. Thus, the prior term becomes $\mathcal{L}_{prior} = ||\theta||^2 + \lambda_u \mathcal{F}_s[\boldsymbol{j}]$, where $\lambda_u$ calibrates the effect of the smoothness in terms of the reprojection uncertainty. This prior term is then combined to obtain $\mathcal{L}_{MAP}$ as described in Section 3.2. We obtain the ensemble prediction by model averaging as in the supervised case (Eq. 3.2). In this case, the data uncertainty for the depth prediction $\widehat{\sigma}^2_{a,m}$ cannot be extracted from the photometric uncertainty parameter $\widehat{u}$. Due to this, only model uncertainty will be considered in the experiments ($\widehat{\sigma}^2_t = \widehat{\sigma}^2_e$).

## 3.5 Teacher-student with uncertain teacher

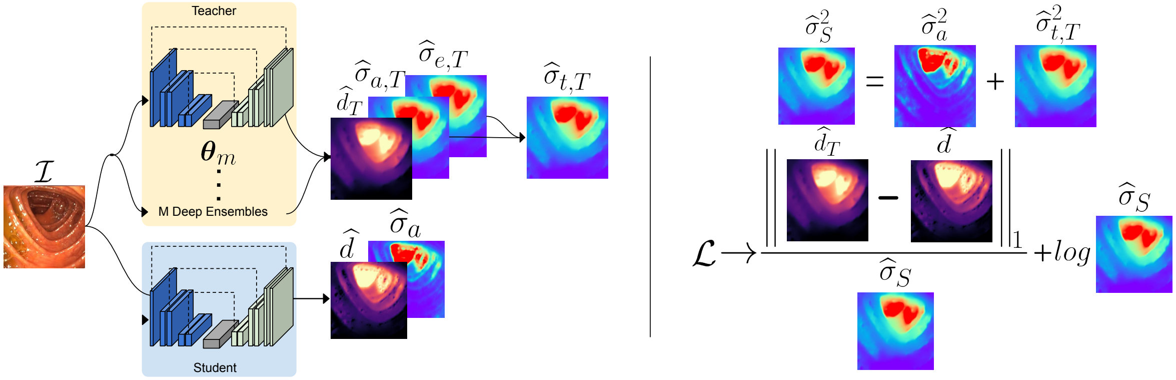


Figure 3.5: Self-supervised training by a teacher-student approach.

In the endoscopic domain, accurate depth training labels can only be obtained from RGB-D endoscopes (which are highly unusual) or synthetic data (that is affected by domain change). We propose the use a of a Bayesian teacher trained on synthetic colonoscopies that produces depth and uncertainty labels. The teacher's epistemic uncertainty allows us to overcome the domain gap automatically. Specifically, our novel teacher-student architecture models depth labels from the predictive posterior of the teacher $d \sim \mathcal{N}(\widehat{d}, \sigma_T^2)$ ($\sigma_T^2$ is the total teacher variance). Thus, the likelihood must incorporate both the teacher and student distributions, which is used in the training loss. As before, the loss is based on a Laplacian likelihood

$$\mathcal{L}_{LL,m} = \frac{1}{w \cdot h} \sum_{\boldsymbol{j} \in \Omega_i} \left( \frac{||\widehat{d}_T[\boldsymbol{j}] - \widehat{d}[\boldsymbol{j}]\,||_1}{\widehat{\sigma}_m[\boldsymbol{j}]} + \log \widehat{\sigma}_m[\boldsymbol{j}] \right) \tag{3.6}$$

where the per-pixel variance is the sum of the teacher predictive variance and the aleatoric one predicted by the student $\widehat{\sigma}_m^2 = \widehat{\sigma}_T^2 + \widehat{\sigma}_{a,m}^2$. Our student is hence aware of the label reliability, which will be affected by the domain change.

## 3.6 Experimental results

We present findings on both simulated and real colonoscopies. Our first dataset is the one generated by [Rau et al., 2019], containing RGB images rendered from a 3D model of the colon in 15 different textures and illumination conditions. The second one, the EndoMapper dataset [Azagra et al., 2022] consists of real monocular colonoscopies.

**Synthetic colon dataset.**

We evaluate three training alternatives: GT (ground truth) depth supervision, SfM supervision and self-supervision. We use 6,550 images for training and 720 images for testing. We observed that training more than 18 networks per ensemble does not improve the performance significantly, so we use this number in our experiments.

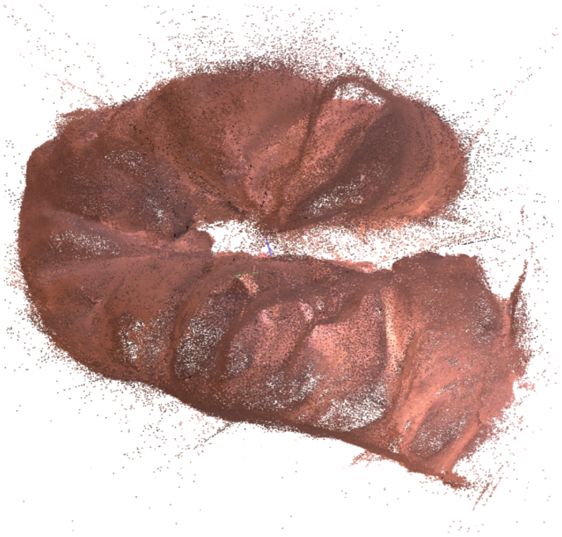


Figure 3.6: Multi-view reconstruction from the colonoscopy synthetic images.

In SfM-related experiments, we use COLMAP [Schonberger and Frahm, 2016]. Figure 3.6 shows the 3D reconstruction from SfM. Since $d^{SfM}$ is up to scale, we compute a scale correction factor $s_i$ per image $\mathcal{I}_i$ as follows:

$$s_i^{SfM} = \frac{\text{median}(d_i^{GT})}{\text{median}(d_i^{SfM})} \tag{3.7}$$

This scale correction is also applied to predictions of self-supervised and supervised SfM models that are also up-to-scale. Table 3.1 shows the metrics for the depth error and its uncertainty.

Supervising a deep ensemble with $d^{GT}$ labels achieves the best depth metrics. In terms of uncertainty, self-supervised and supervised with SfM are underconfident, in contrast to supervised with GT depth, which is overconfident and presents higher (worse) absolute AUCE. Figure 3.7 shows that the aleatoric uncertainty supervised by GT is high around the haustra and in dark areas. The epistemic uncertainty grows with the scene depth. The uncertainty supervised by SfM is high in areas where there are typically holes in SfM reconstructions. Similarly to models trained with GT, the aleatoric uncertainty is also visible in the haustras and the epistemic in the deepest areas. Photometric self-supervision tends to offer the worst performance.

| Approach | $\text{Abs}_{\text{Rel}}$ | $\text{Sq}_{\text{Rel}}$ | RMSE | $\text{RMSE}_{\text{log}}$ | $\delta < 1.25$ | $\delta < 1.25^2$ | $\delta < 1.25^3$ | AUCE |
|---|---|---|---|---|---|---|---|---|
| Supervised GT | **0.050** | **0.335** | **2.996** | **0.102** | **0.978** | **0.993** | **0.997** | +0.190 |
| Supervised SfM | 0.172 | 2.568 | 7.409 | 0.269 | 0.852 | 0.939 | 0.962 | **-0.116** |
| Self-supervised | 0.179 | 1.774 | 7.601 | 0.243 | 0.792 | 0.938 | 0.972 | -0.152 |

Table 3.1: Depth and uncertainty metrics in the synthetic dataset. RMSE in mm.

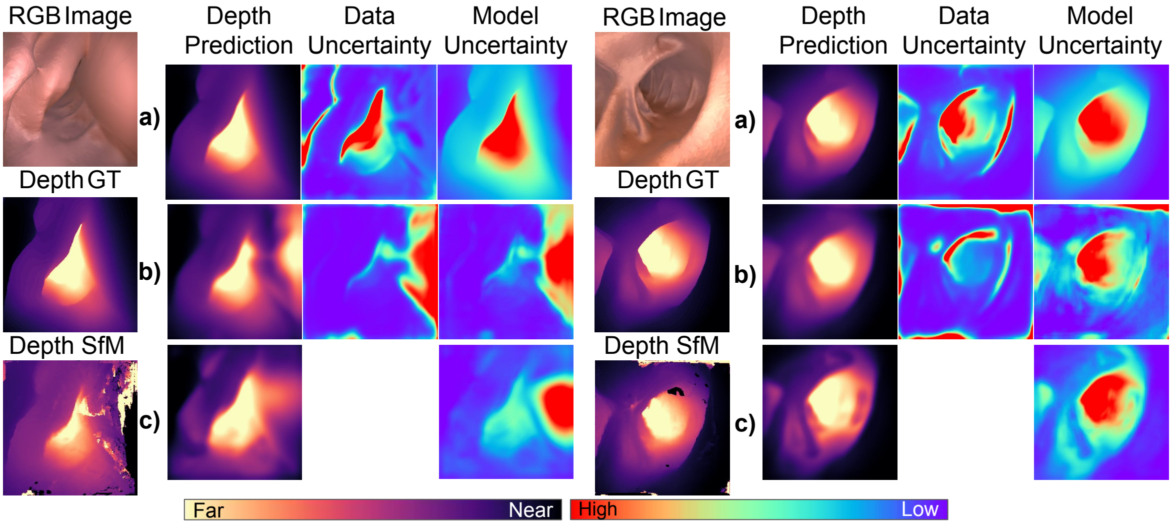


Figure 3.7: Qualitative depth and uncertainty examples of (supervised learning, supervised learning SfM) and self supervised learning in synthetic images. a) Supervised GT, b) Supervised SfM and c) Self-supervised

**EndoMapper dataset.**

This experiment evaluates Bayesian depth networks in real colonoscopies. We use the model previously trained with synthetic ground truth depth ("Supervised GT") to analyse the effect of the domain change. In addition, we also present results from self-supervised training, a baseline teacher-student method [Poggi et al., 2020] and our novel uncertain teacher approach. In real colonoscopies viewpoints change abruptly, images might be saturated or blurry, a considerable amount of liquid might appear and the colon itself produces significant occlusions. For this reasons, we remove images with partial or total visibility issues. We finally use $6{,}912$ images out of the $14{,}400$ images in the complete colonoscopic procedure. In order to obtain depth and uncertainty metrics, we create a 3D reconstruction of the colon using COLMAP (see Figure 3.8). We also use 18-network ensembles for all methods. Table 3.2 shows the results. Our "Uncertain teacher" shows in general the smallest depth errors and the highest correlation between depth errors and predicted uncertainties.

| Approach | $\text{Abs}_{\text{Rel}}$ | $\text{Sq}_{\text{Rel}}$ | RMSE | $\text{RMSE}_{\log}$ | $\delta < 1.25$ | $\delta < 1.25^2$ | $\delta < 1.25^3$ | AUCE |
|---|---|---|---|---|---|---|---|---|
| Supervised GT | 0.240 | 0.644 | 2.595 | 0.308 | 0.645 | 0.898 | 0.962 | -0.148 |
| Self-supervised | 0.371 | 1.260 | 4.603 | 0.431 | 0.417 | 0.721 | 0.886 | -0.273 |
| Teacher-student | 0.234 | 0.600 | 2.532 | 0.301 | 0.657 | 0.903 | 0.963 | -0.328 |
| Uncertain teacher (ours) | **0.230** | **0.572** | **2.458** | **0.298** | **0.667** | **0.906** | **0.964** | **-0.129** |

Table 3.2: Depth and uncertainty metrics in the EndoMapper dataset.

For self-supervised methods, this real setting is challenging due to reflections, fluids and deformations, all of them aspects that are not considered in the photometric reprojection model of self-supervised losses. "Supervised GT" is affected by domain change, as it

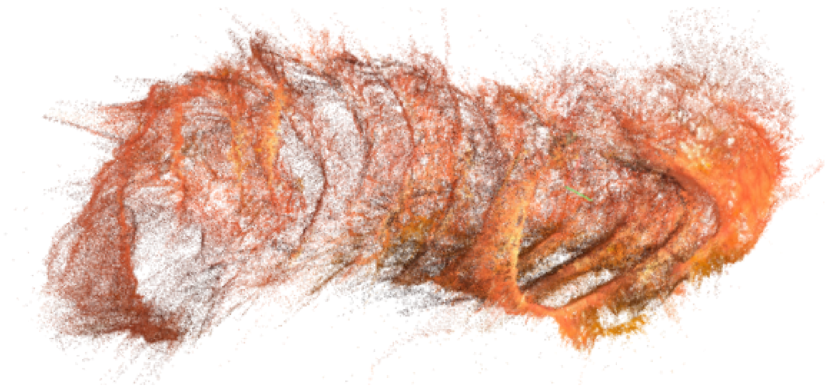


Figure 3.8: Multi-view reconstruction from the real colonoscopy images of the EndoMapper project using COLMAP.

was trained on synthetic data. However, we observe that it successfully generalizes to the real domain and outperforms the self-supervised method. Based on this observation, we use synthetic supervision in the "Teacher-student" baseline and our "Uncertain teacher". In general, teacher-student depth metrics outperform the models trained with GT supervision in the synthetic domain and with self-supervision in the real domain. However, "Teacher-student" presents the worst AUCE metric, as the teacher uncertainty is not taken into account at training time. Our "Uncertain teacher" is the one presenting the best depth and uncertainty metrics, as it appropriately models the noise coming from domain transfer in the depth labels. Figure 3.9 shows qualitative results for the "Supervised GT", "Teacher-student" and "Uncertain teacher" models. Note that the data uncertainty captures light reflection and depth discontinuities in supervised learning. On the other hand, the model uncertainty grows for the deeper areas. Observing these results, we can conclude that the domain change from synthetic to real colon images is not significant. Models trained on synthetic data generalize to real images and outperform models trained with self-supervision on the target domain, due to the challenges mentioned in the previous paragraphs. Finally, figure 3.10 shows a 3D reconstruction based on depth from our method "Uncertain teacher" and the input image.

## 3.7 Conclusions

All systems building on depth predictions from color images benefit from uncertainty estimates, in order to obtain robust, explainable and dependable assistance and decisions. In this chapter, we have explored for the first time supervised and self-supervised approaches for depth and uncertainty single-view predictions in colonoscopies. From our experimental results, we extract several conclusions. Firstly, using ground truth depth as supervisory signal outperforms self-supervised learning and results in better calibrated models. Secondly, approaches based on photometric self-supervision and on SfM supervision coexist in the

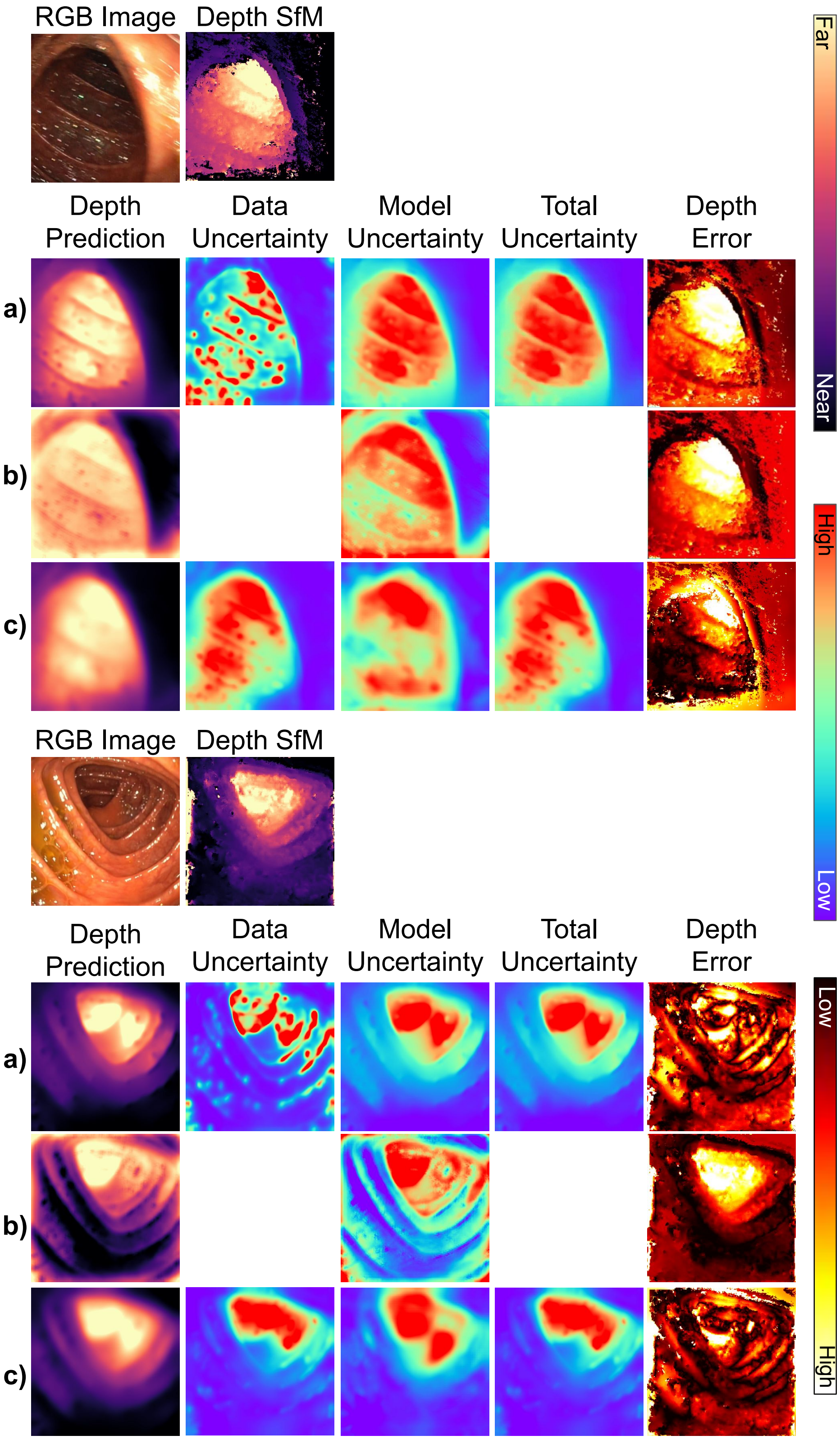


Figure 3.9: Qualitative depth and uncertainty examples for a) supervised, b) self-supervised, and c) uncertain teacher-student learning in real images.

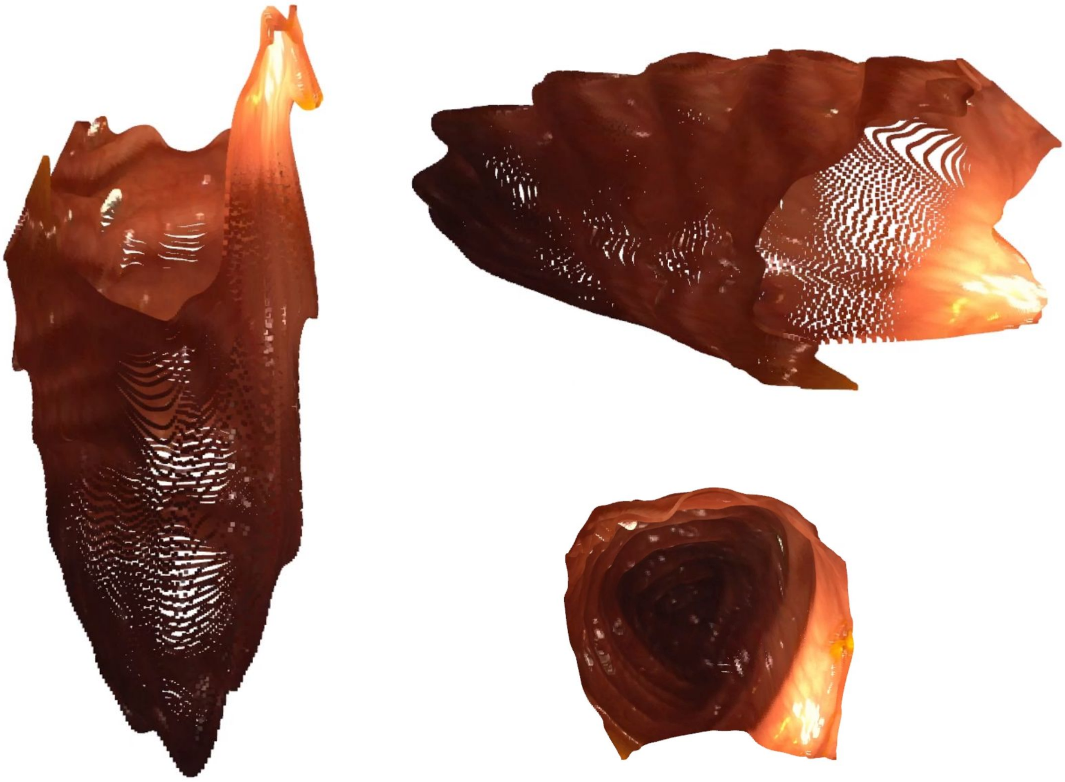


Figure 3.10: 3D reconstruction based on input image and depth from our Uncertain teacher method.

literature and there is a lack of analysis and results showing which type is more convenient. Thirdly, our experiments show that models trained in synthetic colonoscopies generalize to real colonoscopy images. Finally, we have proposed a novel teacher-student architecture that incorporates the teacher uncertainty in the loss, and have shown that it produces lower depth errors and better calibrated uncertainties than previous teacher-student architectures.

# Chapter 4

# LightDepth: single-view depth self-supervision from illumination decline

In this chapter, we introduce a novel self-supervision methodology that exploits illumination decline data of systems where source light and cameras are co-located, paving the way for enhanced depth estimation in endoscopic settings. Several works have proposed methods based on supervised learning, which struggle with simulation to real domain shifts. On the other hand, methods based on multi-view self-supervision face challenges related to the specific endoscopic domain, such as deformations and small baselines. Assuming both limitations, we approach the problem from a different perspective by including a light model in the method. Our experiments show that by doing that we are able to outperform both limitations in the literature, outperforming all previous approaches.

## 4.1 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.

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.

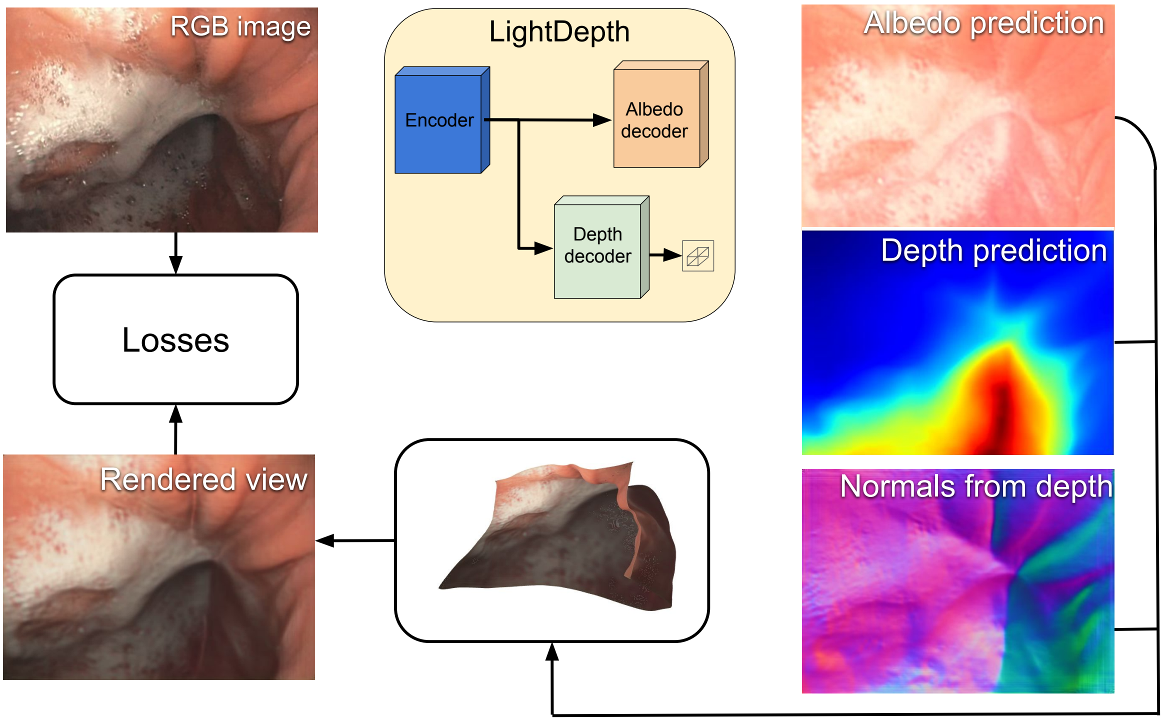


Figure 4.1: **Single-view depth self-supervision in LightDepth.** A two-headed deep network predicts albedo and depth from a single image and estimates surface normals from predicted depths. These are used to render a new image, that takes into account illumination decline and the endoscope's photometric 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.

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 4.1, 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 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 fully 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 that can be trained in large colonoscopy datasets without requiring ground truth labels and even further refined online in the test views.

## 4.2 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 the work of Laina et al. [2016] and more recently to transformers [Ranftl et al., 2021, Bhat et al., 2021, Li et al., 2022]. Some of those works [Bhat et al., 2021, Li et al., 2022] also discretize the continuous depth space into bins and formulate the problem as an ordinal regression, as done by Fu et al. [2018]. Other advances include interpretability [Dijk and de Croon, 2019], uncertainty quantification [Poggi et al., 2020, Rodríguez-Puigvert et al., 2022], and modeling camera intrinsics [Facil et al., 2019, Gordon et al., 2019]. All these 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, Johnston and Carneiro, 2020, Godard et al., 2019]. 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 [Karaoglu et al., 2021, Freedman et al., 2020, 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, a conditional GAN is used for depth recovery while integrating SLAM and multi-view inputs by Shen et al. [2019]. Chen et al. [2019b] trained a depth network 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., 2020] 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 by 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 shows that this still holds when the light source is translated with respect to the optical centre is provided in [Collins and Bartoli, 2012a, 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 [Hao et al., 2020a, Parot et al., 2013, Collins and Bartoli, 2012b].

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, Sengupta et al., 2018], 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 endoscope calibration [Modrzejewski et al., 2020, Hao et al., 2020b, Azagra et al., 2022, Batlle et al., 2022].

## 4.3 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 in a differentiable rendering pipeline (Figure 4.2). 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.

### 4.3.1 Photometric model

As in the work of Batlle et al. [2022] and Modrzejewski et al. [2020], 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 Figure 4.3. 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 $\psi_i$, we write its radiance as

$$\sigma_{SLS}(\mathbf{x}_i, \psi_i) = \frac{\sigma_0}{\|\mathbf{x}_i - \mathbf{x}_l\|^2} R(\psi_i) \tag{4.1}$$

$$R(\psi_i) = e^{-\mu(1-\cos(\psi_i))} \tag{4.2}$$

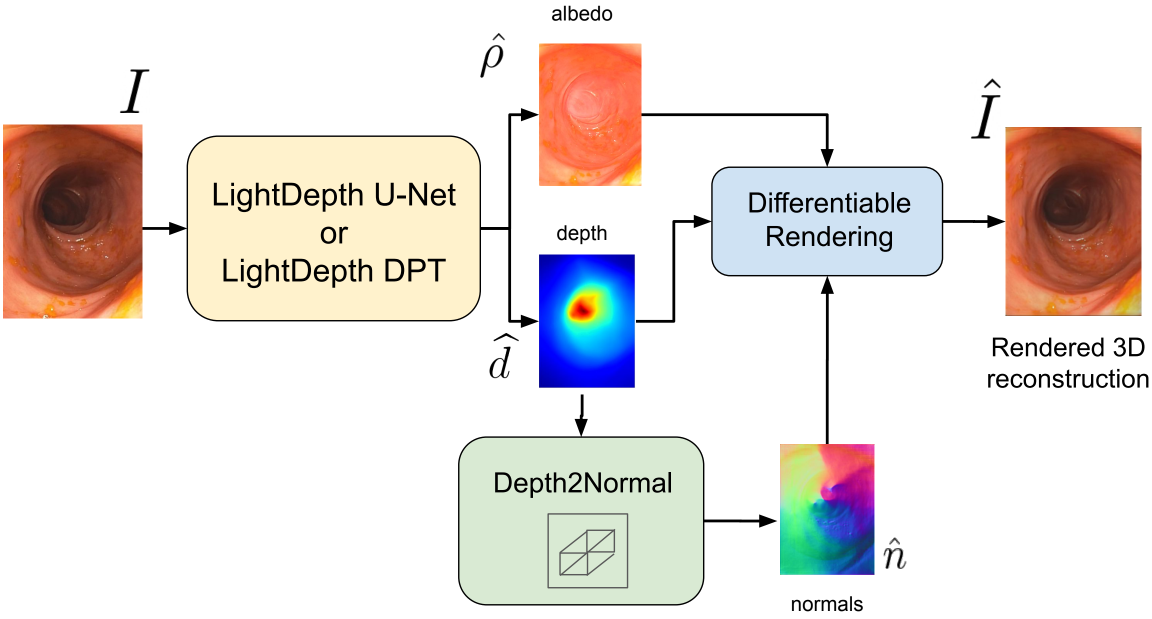


Figure 4.2: **Differentiable Rendering:** The input image is fed into a neural network that predicts albedo and depth values for each pixel. From the estimated depths, we compute the normals at each pixel surface using a kernel-based approach. Then, the depths, albedos, and normals are sent to a differentiable 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.

where $\sigma_0$ is the maximum radiance and $R(\psi_i)$ is the radial attenuation controlled by a spread factor $\mu$. 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 $\mathbf{x}_i$.

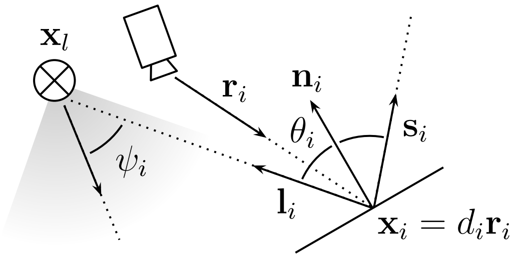


Figure 4.3: Spotlight illumination model, a spotlight source at position $\mathbf{x}_l$ illuminates the surface point $\mathbf{x}_i$. The emission has $R(\psi_i)$ radial fall-off, suffers from an inverse-square decline with $\mathbf{x}_l \rightarrow \mathbf{x}_i$ and attenuates with the incidence angle ($\theta_i$). $\mathbf{l}_i$, $\mathbf{n}_i$, $\mathbf{r}_i$ and $\mathbf{s}_i$ are unit vectors.

### 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

$$\mathcal{I}(d_i, \rho_i, g) = \left( \frac{\sigma_0}{\|d_i\mathbf{r}_i - \mathbf{x}_l\|^2} R(\psi_i) \cos(\theta_i)\ \rho_i\ g \right)^{1/\gamma}, \tag{4.3}$$

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. $\theta_i$ stands for the light's incidence angle with respect to the surface normal $\mathbf{n}_i$, such that, $\cos\theta_i = \mathbf{l}_i \cdot \mathbf{n}_i$. $\rho_i$ represents the albedo of the surface at that point. $g$ denotes the gain applied by the camera and $\gamma$ 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 4.3.2.

### Calibration.

Each endoscope has different geometric and photometric parameters, the former affecting the inverse project model $\pi^{-1}$ 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 [Azagra et al., 2022, Batlle et al., 2022]. 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.

### 4.3.2 Self-supervision losses

Formally, the network of Figure 4.2 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 $\rho$ 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

$$\mathcal{L} = \mathcal{L}_p + \lambda_s \mathcal{L}_s + \lambda_{sp} \mathcal{L}_{sp} \ , \tag{4.4}$$

where $\lambda_s$ and $\lambda_{sp}$ 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 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.

$$\mathcal{L}_p = \sum_{i \in \Omega} (I_i - \widehat{I}_i)^2, \quad \text{where} \quad \widehat{I}_i = \mathcal{I}(i, \widehat{d}_i, \widehat{\rho}_i, g) \tag{4.5}$$

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

$$\mathcal{L}_s = |\partial_x \widehat{d}| e^{-|\partial_x I|} + |\partial_y \widehat{d}| e^{-|\partial_y I|} \tag{4.6}$$

Finally, recall that we made a Lambertian assumption in Eq. 4.3, 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

$$\mathbf{s}_i = \mathbf{l}_i - 2\hat{\mathbf{n}}_i \left( \hat{\mathbf{n}}_i \cdot \mathbf{l}_i \right) \tag{4.7}$$

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

$$\mathcal{L}_{sp} = \sum_{i \in \Omega} \left( m_i \left( \mathbf{s}_i \cdot (-\mathbf{r}_i) - 1 \right) \right)^2 \ , \tag{4.8}$$

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

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. 4.3, 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.

### 4.3.3 Network architecture

Our network outputs depth and albedo maps. In Figure 4.4 and Figure 4.5, 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 [He et al., 2016] serving as the backbone, initialized with the weights from ImageNet [Deng et al., 2009]. Our decoders design is inspired by  [Godard et al., 2019], our albedo decoder uses sigmoid activation function and our depth decoder ELU+1 activation function after the last convolution. The second one relies on visual transformers for depth estimation [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, upsampling the outputs with bilinear interpolation. As a backbone, we use a Resnet-50 (DPT-Hybrid) and two decoders that reassemble the tokens and apply attention heads.

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.

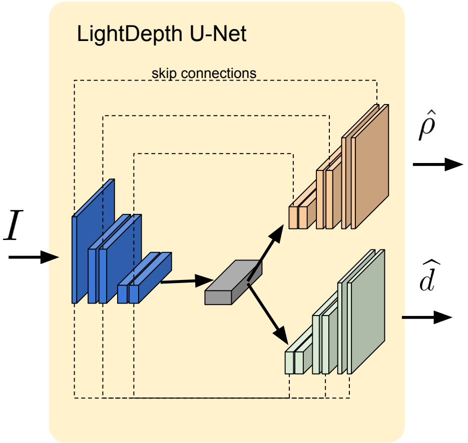


Figure 4.4: LightDepth U-Net is based on a standard U-Net [Ronneberger et al., 2015] with two decoding branches.

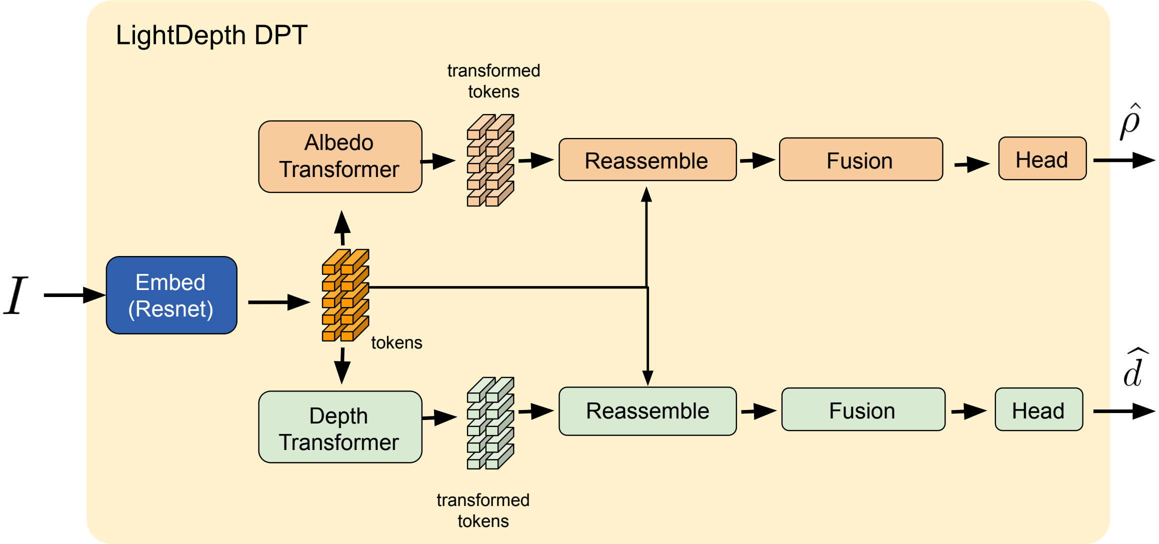


Figure 4.5: LightDepth DPT is based on the DPT-Hybrid architecture [Ranftl et al., 2021], with a second decoder branch added for the albedo.

## 4.4 Results

### 4.4.1 Datasets

We evaluated LightDepth and relevant baselines on three endoscopy datasets: An in-house *synthetic colon*, *C3VD* [Bobrow et al., 2022], and *EndoMapper* [Azagra et al., 2022]. 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. [2022]. 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., 2022]. 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. Figure 4.6 depicts samples from the synthetic dataset, which contains ground truth data on depth, albedo, and normals.

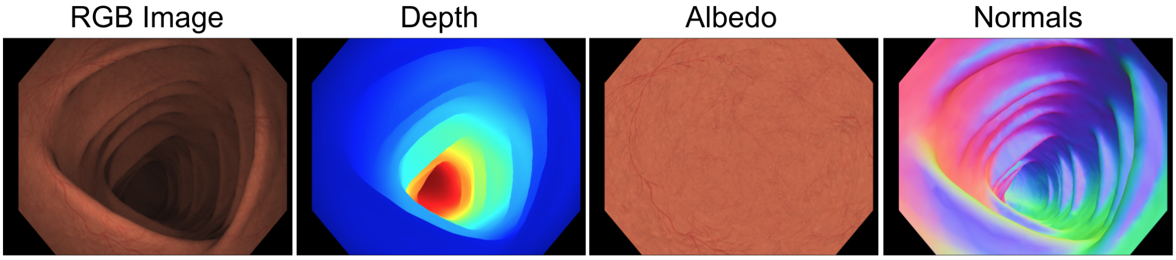


Figure 4.6: Samples from in-house Fish-eye lambertial synthetic dataset.

**Colonoscopy 3D Video Dataset**

**C3VD** [Bobrow et al., 2022] 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 the Scaramuzza model [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. Table 4.1 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.

| 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 4.1: Dataset Split for C3VD

### EndoMapper

**EndoMapper** [Azagra et al., 2022] provides the most challenging data, as it contains 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.

### 4.4.2 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. 4.4 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 4.4, 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 4.3 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 4.8 and Figure 4.9 show 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.

### 4.4.3 Quantitative and qualitative results on synthetic and phantom

#### Synthetic colon.

In Table 4.2, we 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

|  |  |  |  |  |  |  | Depth [mm] |  |  |  |  | Normals [°] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Architecture | Backbone | Supervision | MAE ↓ | MedAE ↓ | RMSE ↓ | $\text{RMSE}_{\text{log}}$ ↓ | $\text{Abs}_{\text{Rel}}$ ↓ | $\text{Sq}_{\text{Rel}}$ ↓ | $\delta < 1.25$ ↑ | $\delta < 1.25^2$ ↑ | $\delta < 1.25^3$ ↑ | MAE ↓ |
| 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** |

Table 4.2: Depth and normal metrics for several architectures and supervision modes. Best results per dataset are boldfaced, second best underlined.

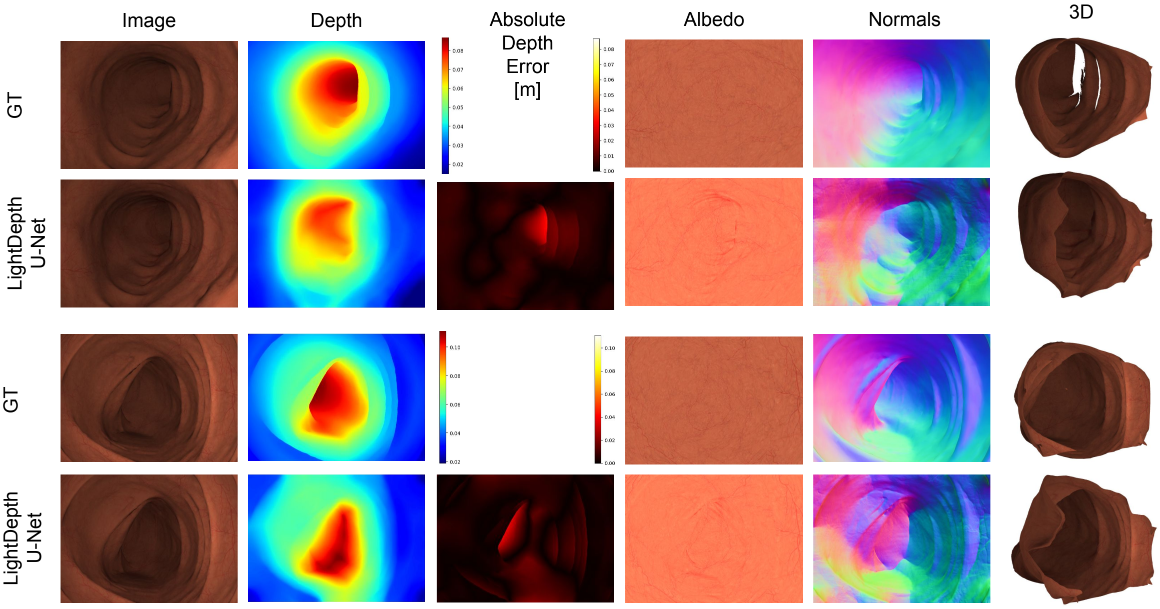


Figure 4.7: Qualitative examples of LightDepth in Synthetic dataset.

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.

|  |  |  |  |  |  |  | Depth [mm] |  |  |  |  | Normals [°] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Architecture | Backbone | Supervision | MAE ↓ | MedAE ↓ | RMSE ↓ | $\text{RMSE}_{\text{log}}$ ↓ | $\text{Abs}_{\text{Rel}}$ ↓ | $\text{Sq}_{\text{Rel}}$ ↓ | $\delta < 1.25$ ↑ | $\delta < 1.25^2$ ↑ | $\delta < 1.25^3$ ↑ | MAE ↓ |
| 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 | ResNet50 | Depth GT | **3.22** | 2.77 | **4.10** | **0.0860** | **0.0699** | **0.0004** | **0.9640** | 0.9865 | 0.9913 | **15.1** |
| Monodepth2 | ResNet50 | Multi-View | 14.27 | 9.59 | 18.64 | 0.3921 | 0.2971 | 0.0070 | 0.4897 | 0.7313 | 0.8611 | 43.6 |
| CADepth | ResNet18 | Multi-View | 52.35 | 17.04 | 87.43 | 0.9144 | 1.1916 | 0.2650 | 0.3664 | 0.5653 | 0.6679 | 67.2 |
| XDCycleGAN | 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 4.3: Depth and normal metrics for several architectures and supervision modes. Best results per dataset are boldfaced, second best underlined.

We report depth and normal metrics on the real phantom images of the C3VD dataset in

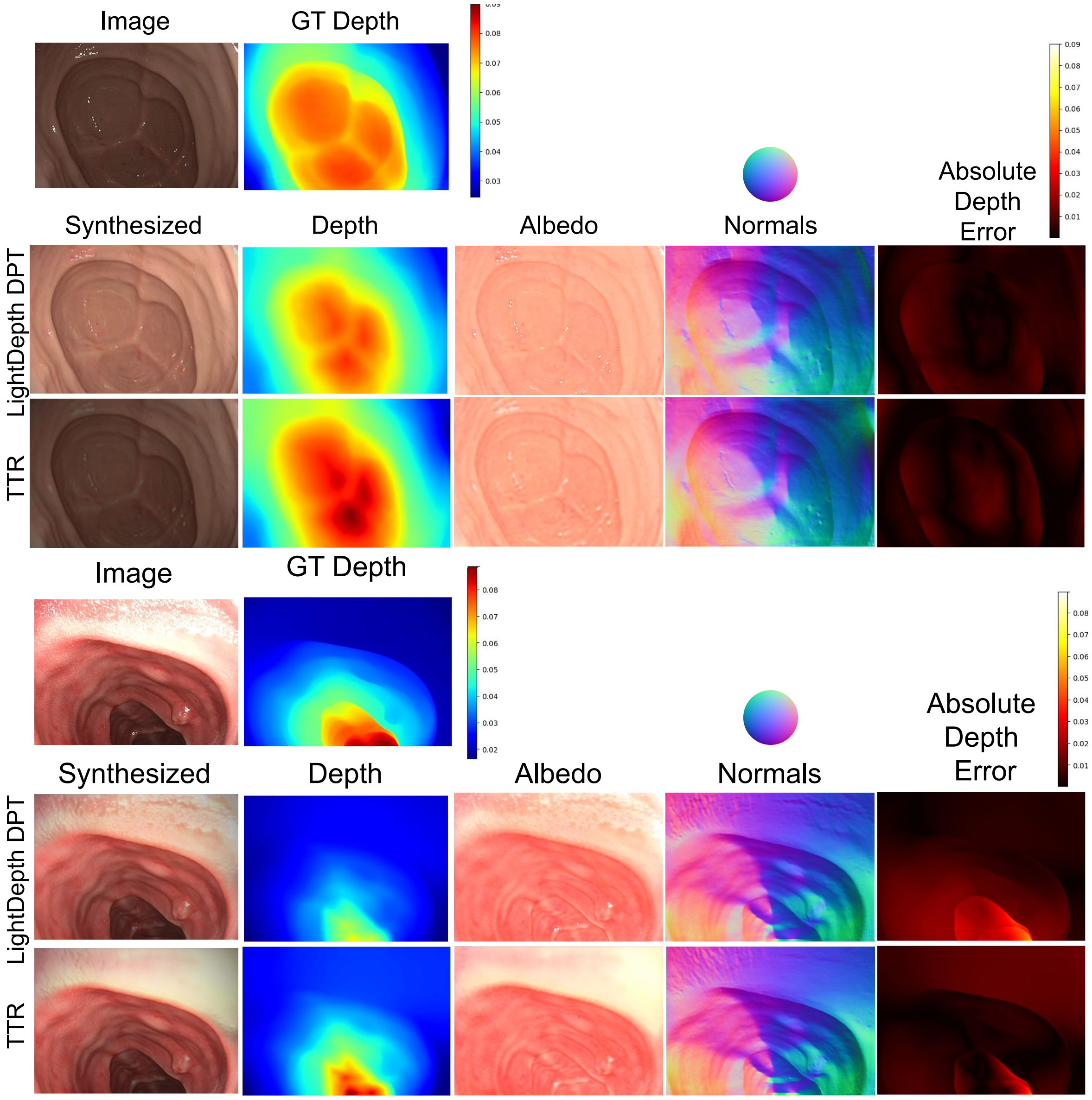


Figure 4.8: LightDepth and LightDepth TTR on C3VD. Our light decline captures the correct shape of the cecum in the first image and the shape of the polyp in the second. Note how the estimates of normals and albedo are similar before and after TTR. By optimising depth by reducing illumination, DepthLight achieves a darker appearance and improvements in depth estimation.

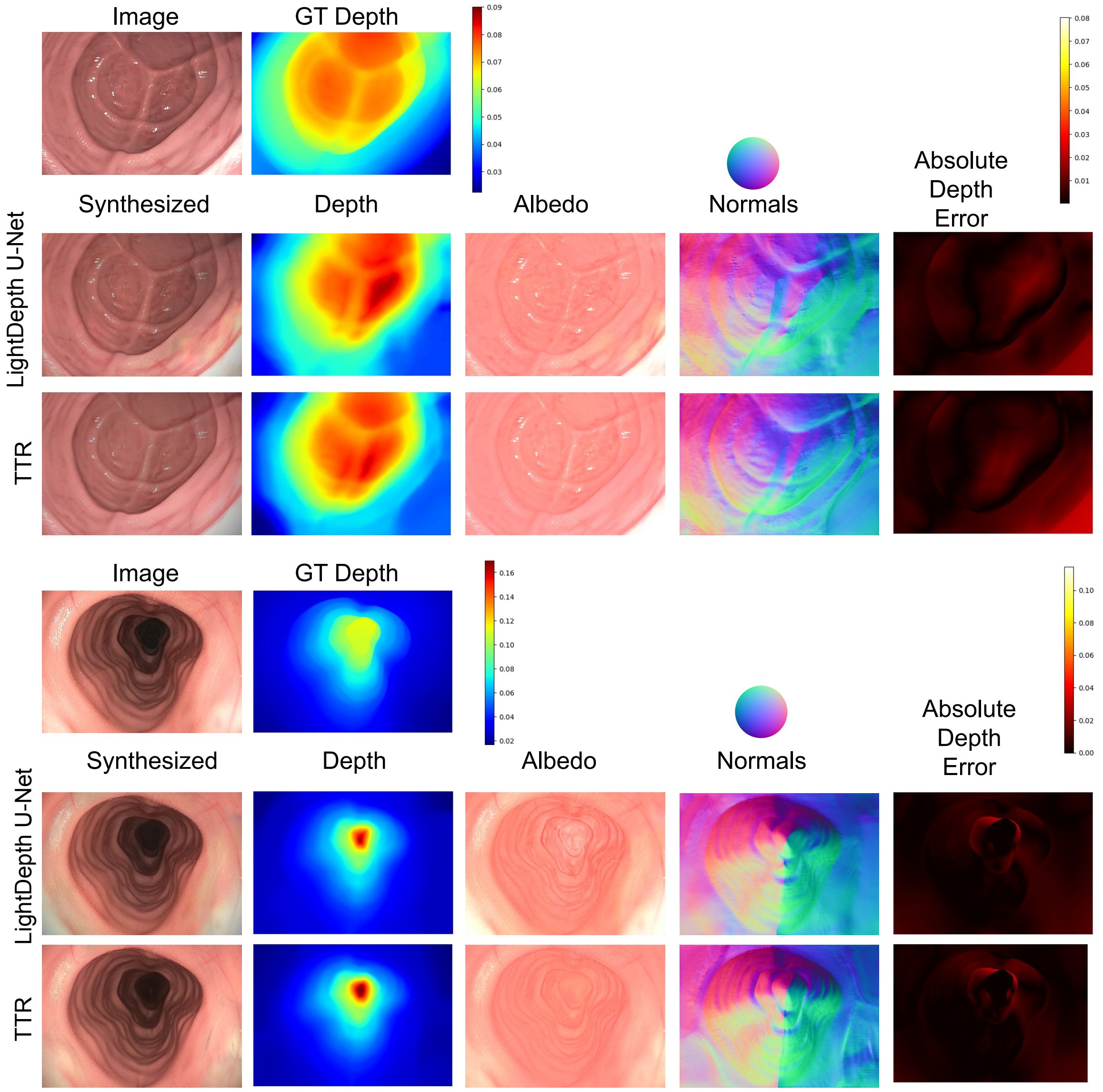


Figure 4.9: LightDepth and LightDepth TTR on C3VD. Quantitative examples of LightDepth U-Net in C3VD.

Table 4.3. 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 4.4 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 4.3, 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 the phantom colon, 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.

| 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 4.4: SSIM and MAE for rendered images in C3VD. Test-time refinement (TTR) gives a substantial improvement.

### 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., 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 4.5 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

| 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) | 3.80 | **2.51** | 5.54 | 23.6 |
| C3VD | C3VD | Depth GT | 4.15 | 3.29 | 5.52 | 26.5 |
| | | Light | 4.37 | 2.92 | 6.31 | 24.0 |
| | | Light (TTR) | **3.72** | 2.59 | **5.43** | **23.5** |

Table 4.5: 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.

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.

Figure 4.11 shows examples of Open3d [Zhou et al., 2018b], in-house, U-Net and TFtN [Fan et al., 2021] used in the analysis.

### Normals from depth

The literature details different manners to obtain surface normals from a depth map, e.g., [Fan et al., 2021, Boulch and Marlet, 2016]. Table 4.10 shows a MAE analysis of the most promising ones in C3VD. Specifically, we evaluate 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. Figure 4.11 shows examples of Open3d [Zhou et al., 2018b], in-house, U-Net and TFtN [Fan et al., 2021] used in the analysis. 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.

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

Figure 4.10: Normal's MAE for baseline methods.

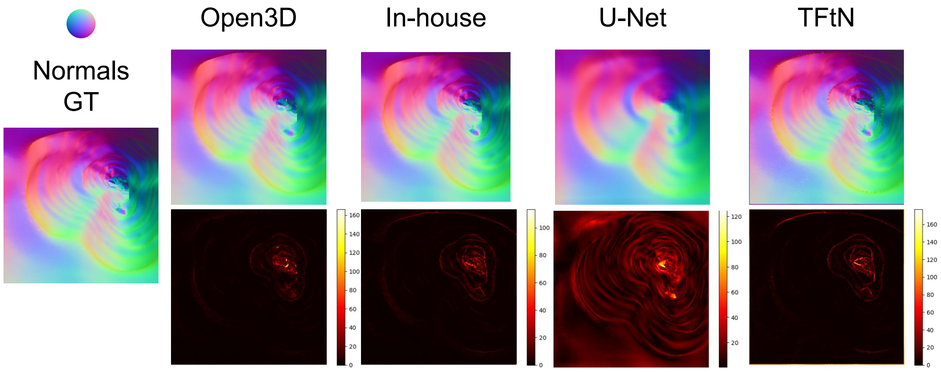


Figure 4.11: Quantitative results of different approaches to obtaining surface normals from a depth map.

### Ablation study on the loss

In Table 4.6, 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.

## 4.4.4 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 4.12, Figure4.13 and Figure 4.14. Some details are recovered very accurately, such as the normal maps showing clearly the tubular shape; the depth maps reflecting the discontinuities in the Haustras; the albedos capturing the blood vessels, in particular in the 5th column; and the bubbles and fluids colors in the 6th and 7th 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

| | Depth [mm] | | | Color | Normals [°] |
|---|---|---|---|---|---|
| Loss | MAE ↓ | MedAE ↓ | RMSE↓ | MAE ↓ | 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 4.6: Ablation study of the losses with LightDepth U-Net in C3VD dataset. Observe the improvement given by each term.

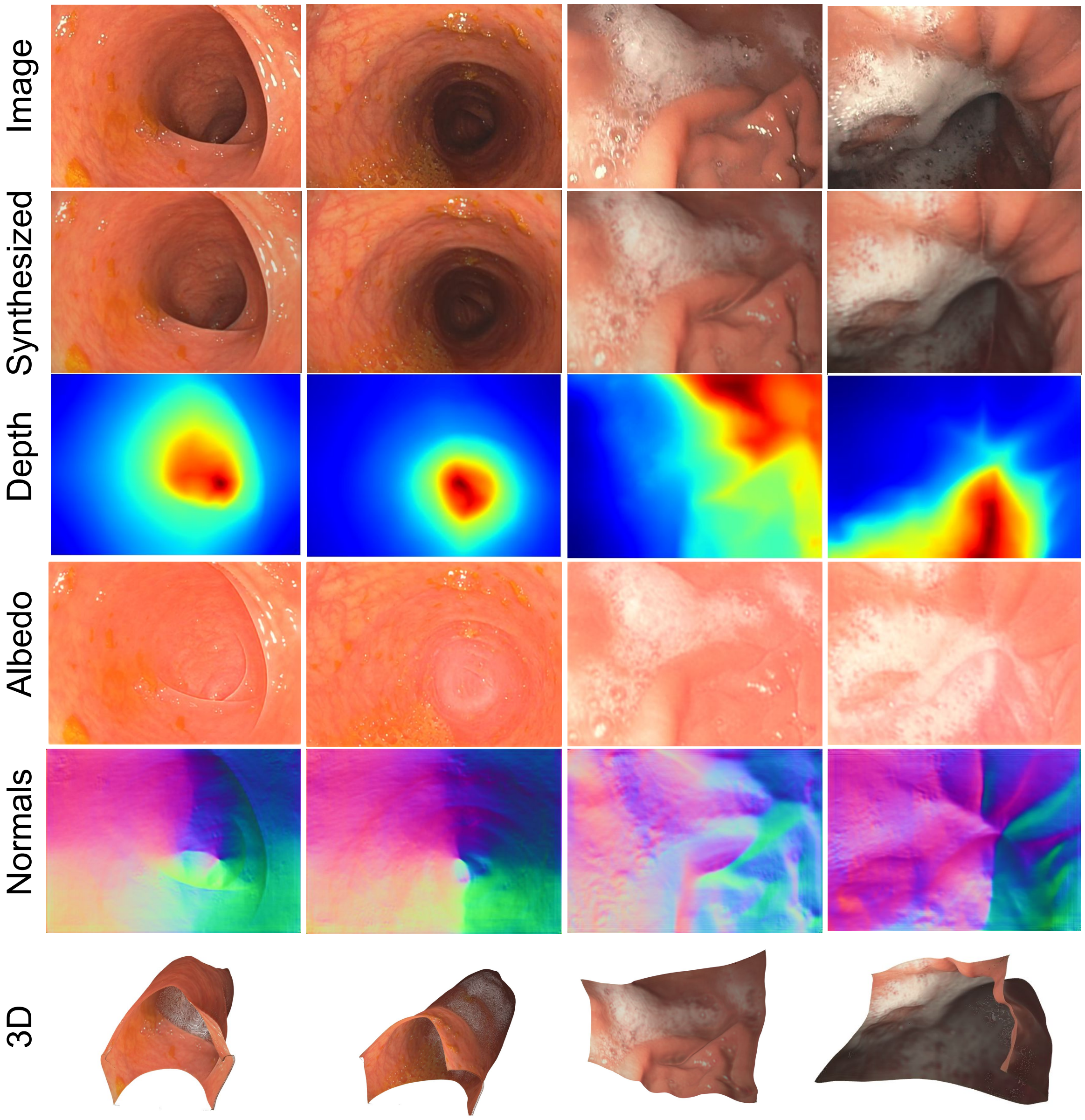


Figure 4.12: Qualitative results on EndoMapper with LightDepth DPT. In colonoscopies, observe that the normals exhibit a tubular shape specific of the colon. The albedo prediction captures disruptions such as veins, dirt, foam and specularites. Note the influence of light decline in the image and the correlation with the estimated depths.

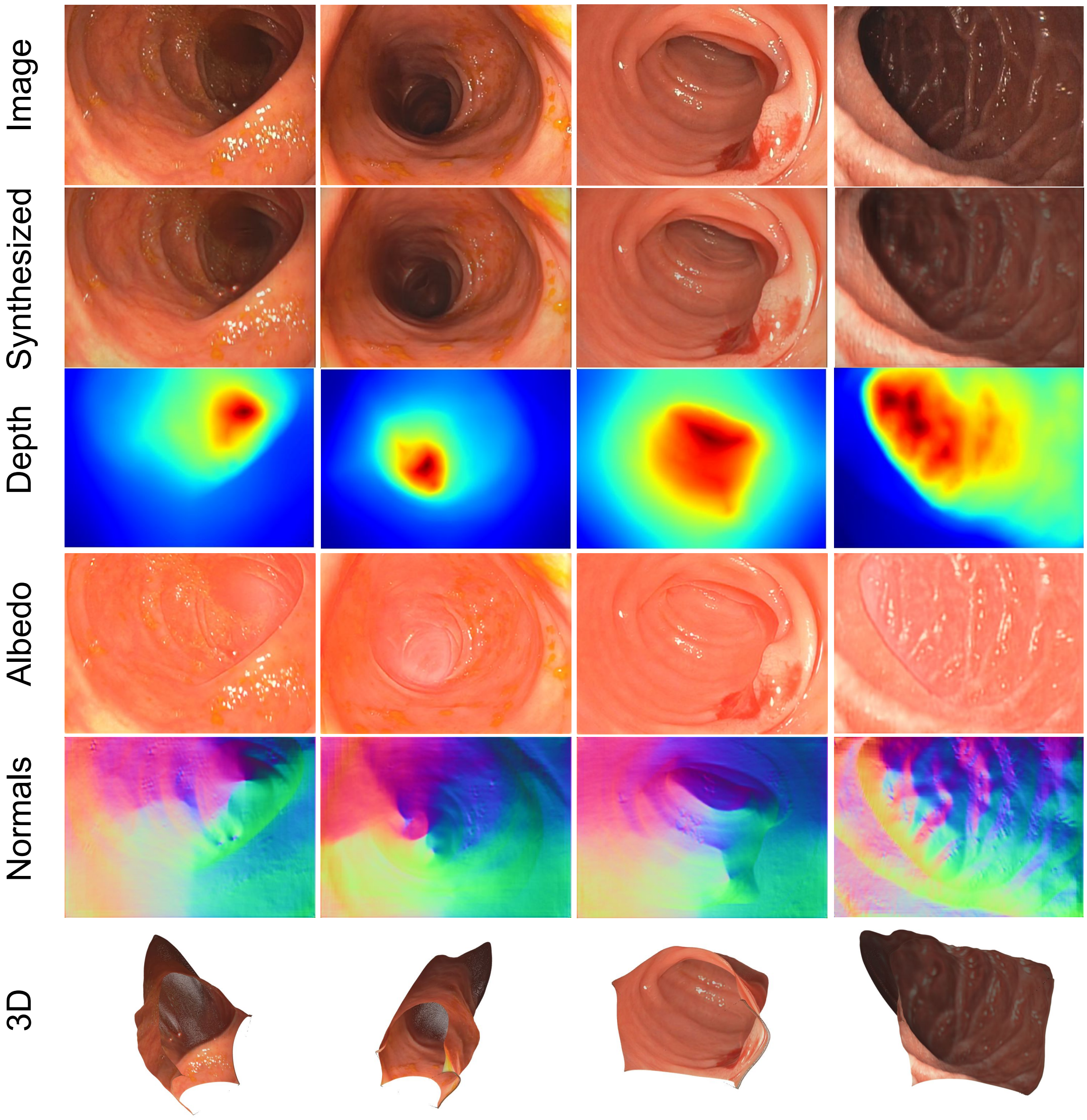


Figure 4.13: Qualitative results on EndoMapper with LightDepth DPT. The albedo prediction captures blood. Note the influence of light decline in the image and the correlation with the estimated depths.

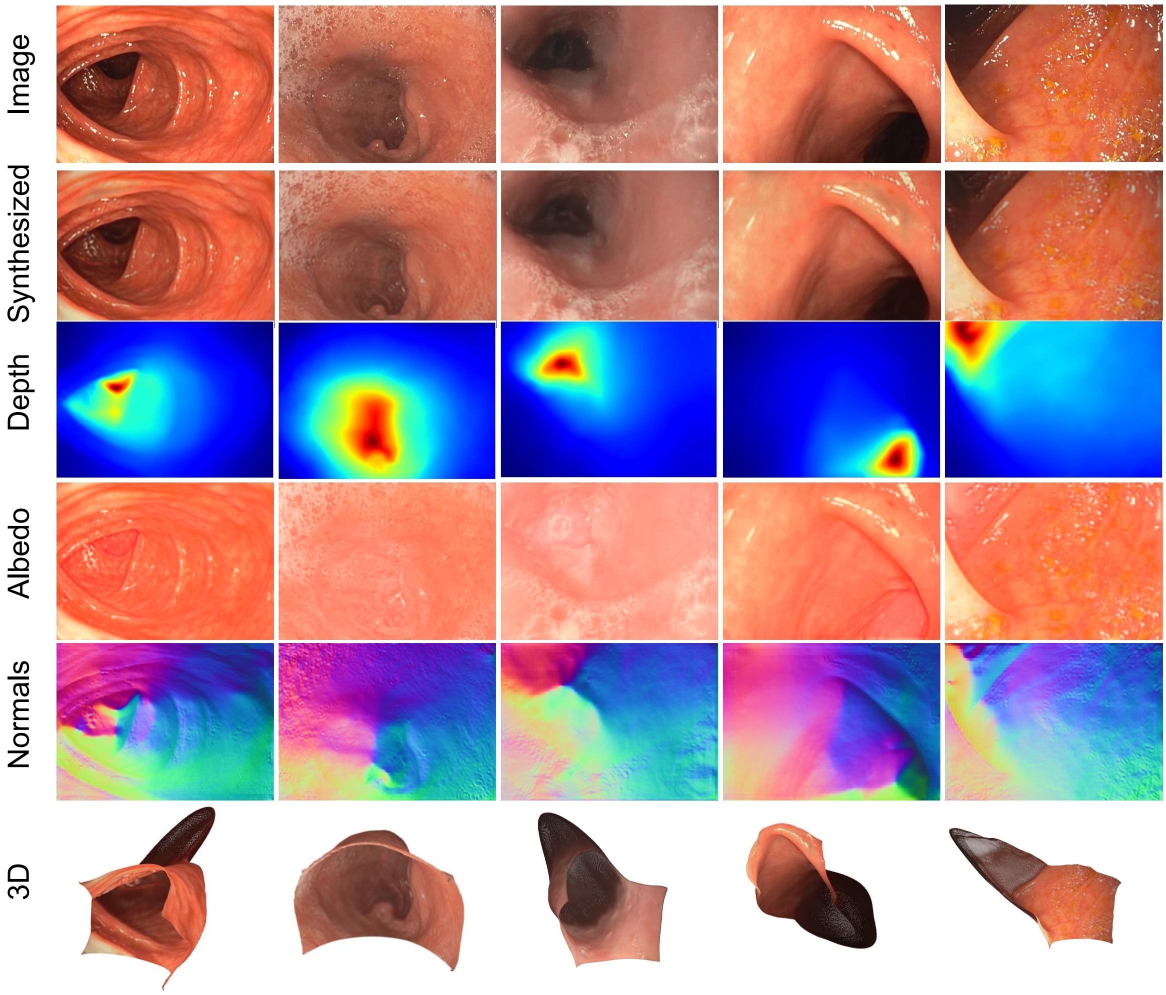


Figure 4.14: Qualitative results on EndoMapper with LightDepth DPT.

results hints that the strengths of our techniques demonstrated quantitatively in Section 4.4.3 will carry over on truly realistic scenarios like this one.

## 4.5 Limitations and discussion

As mentioned in Section 4.3.2, our depth predictions are up-to-scale. Even if the camera auto-gain were 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.

## 4.6 Conclusions

In this chapter, 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.

# Chapter 5

# Conclusions and Future work

In this thesis, we investigate several aspects related to depth estimation from a single view, presenting a number of contributions. Specifically, our contributions focus on self-supervision and uncertainty quantification.

In the work described in the first chapter, we have made several findings related to uncertainty quantification with Bayesian deep neural networks for supervised learning of single-view depth. Among the most relevant it is the insight that the specific layers in which MC dropout is applied in a Bayesian neural architecture plays a critical role in determining the accuracy of both depth and uncertainty estimates. In our experiments, deep ensembles stand out as the method with the best uncertainty calibration. However, the application of MC dropout, when done in the encoder, manifested a performance that outperformed other variants highlighted in the literature. We observed that aleatoric uncertainty is relatively straightforward to capture. Nonetheless, it is important to highlight the relevance of epistemic uncertainty, which our research identified as a significant part of the total uncertainty. Current scalable methods still have room for improvement, and achieving a reasonable calibration of the uncertainty remains an open goal.

In Chapter 3, on the uncertain single-view depths in colonoscopies, we conduct research on single-view depth uncertainty quantification, in this case with focus in medical colonoscopy images, bringing a novel perspective to the established literature. Our main motivation is highlighting the importance of both accurate depth and uncertainty estimations in medical imaging. Specifically, in colonoscopy images, depth uncertainty is most prevalent in specular reflections and overexposed regions, dark areas where the color signal is weaker and depth discontinuities such as haustra. We demonstrate empirically that the uncertainty captured by deep ensembles behaved coherently when a domain change occurred. This finding has a high relevance, as in the endoscopic domain it is very difficult or impossible to have ground truth annotations, and hence simulation-to-real domain transfer is common. Our research also led to the development of a novel self-supervised method based on an uncertainty-aware teacher-student architecture. Our novel architecture incorporates the teacher uncertainty into the learning pipeline. Our experiments show a notable efficiency in reducing depth errors and improving the uncertainty calibration compared to its predecessors. While our method represents a promising advance, its effectiveness is largely dependent on the skill of the teacher model. New techniques need to be developed not only to capture uncertainty, but also to be able to apply it to deep learning.

Finally, in the fourth chapter, we propose a new single-view self-supervised learning that models quadratic light attenuation and unlocks single-view self-supervised learning in natural medical images. Our so-called LightDepth method offers an original approach to depth self-supervision. It enables consistent depth estimation without the need for ground truth supervision, based on the use of illumination as a supervisory signal for depth. Our research shows that the illumination-decline self-supervision is significantly superior to multi-view self-supervision, and even matches ground truth supervision. Moreover, and as a result of using just a single view for self-supervision, a distinct advantage of our method is its ability to refine on-the-fly during testing, which further improves the depth predictions.

Our results indicate that while supervised approaches can often be effective for endoscopic vision, their performance is also limited by the particularities of the use case, specifically the lack of depth sensors in minimally invasive procedures and hence the difficulty of capturing data with depth annotations. We also observed that multi-view self-supervision is limited by the difficulties of estimating relative motion from colonoscopy images, which is also a research challenge in itself. We believe that our results open a promising path, as our single-view self-supervision overcomes the limitations of both approaches and does only requires RGB data, which does not pose serious difficulties for being acquired .

# Conclusiones y trabajo futuro

En esta tesis, se investiga varios aspectos relacionados con la estimación de profundidad desde una única vista. Las contribuciones se enfocan en sistemas de aprendizaje autosupervisado y en la cuantificación de incertidumbre en redes neuronales profundas. En el primer capítulo, hemos aportado varios descubrimientos relativos a la cuantificacion de incertidumbre para la estimación de profundidad con redes bayesianas profundas en un aprendizaje supervisado utilizando una única vista. Entre los más relevantes se encuentra la observación de que la aplicación de MC dropout en diferentes capas de la arquitectura de la red tiene un papel determinante respecto a la precisión de la estimación de profundidad e incertidumbre. En nuestros experimentos, deep ensembles destaca como el método con mejor calibración de incertidumbre. Sin embargo, la aplicación de MC dropout, cuando se hace en el encoder, ha demostrado un rendimiento mayor que otras variantes de la literatura. Hemos observado que la incertidumbre aleatórica es razonablemente fácil de capturar. No obstante, es importante destacar la relevancia de la incertidumbre epistémica como una parte significativa de la incertidumbre total. Los métodos escalables tienen todavía potencial de mejora, dado que, lograr una calibración perfecta de la incertidumbre sigue siendo un objetivo abierto.

En el Capítulo 3, realizamos una investigación sobre la cuantificación de la incertidumbre respecto a la estimación de profundidad. En este caso, nos enfocamos en las imágenes médicas de colonoscopia, aportando una perspectiva novedosa a la literatura establecida.

Nuestra motivación es resaltar la importancia de estimaciones precisas y calibradas con respecto a la profundidad y la incertidumbre en las imágenes médicas. En concreto, en las imágenes de colonoscopia, la incertidumbre de profundidad es más pronunciada en los reflejos especulares y las regiones sobreexpuestas, en zonas oscuras donde la señal de color es más débil y también las discontinuidades de profundidad, como las haustras. Se demuestra empíricamente que la incertidumbre capturada por deep ensembles se comporta de forma coherente cuando se produce un cambio de dominio. Este hallazgo tiene una gran relevancia, ya que en el dominio endoscópico es muy difícil disponer de anotaciones ground truth y, por lo tanto, la transferencia de dominio de simulación a real es una práctica común.

En nuestra investigación proponemos un novedoso método autosupervisado basado en una arquitectura teacher-student que incorpora la incertidumbre del teacher en el proceso de aprendizaje. Nuestros experimentos muestran que existe una notable eficacia a la hora de reducir los errores de profundidad y mejorar la calibración de la incertidumbre en comparación con otros métdos de aprendizaje. Nuestro método representa un avance prometedor, aunque su eficacia depende en gran medida de la habilidad del teacher. Por último, en el cuarto Capítulo, proponemos un nuevo aprendizaje autosupervisado desde una única vista, el cual modela la atenuación cuadrática de la luz y desbloquea el aprendizaje autosupervisado de una sola vista para imágenes médicas naturales. Nuestro método denominado LightDepth ofrece un nuevo enfoque para la autosupervisión de la profundidad, ya que permite una estimación de profundidad sin la necesidad de supervisión con ground truth, basándose en el uso de la iluminación como señal supervisora. Nuestra investigación demuestra que la autosupervisión utilizando la información de la iluminación es significativamente superior a la autosupervisión multivista, e incluso iguala a los métodos de supervisión con ground truth. Dado que el aprendizaje propuesto requiere únicamente el uso de una sola vista para el entrenamiento, hace que nuestro método sea capaz de refinar sus predicciones con una optimización local, lo que mejora aún más las predicciones de profundidad.

Nuestros resultados indican que, aunque los métodos supervisados pueden ser eficaces para la visión endoscópica, su rendimiento se ve limitado por las particularidades del entorno. La falta de sensores de profundidad en los procedimientos mínimamente invasivos supone una dificultad a la hora de capturar datos con anotaciones de profundidad ground truth, valiosos para el entrenamiento supervisado. Finalmente, observamos que la autosupervisión multivista se ve limitada por las dificultades para estimar el movimiento relativo de la cámara entre imágenes, lo que también constituye un reto de investigación en sí mismo. Creemos que nuestros resultados abren un camino prometedor, la autosupervisión de una sola vista supera las limitaciones de ambos enfoques y sólo requiere datos RGB, que no plantean dificultades para ser adquiridos.

# Bibliography

Amit Aides, Ariel Gordon, Daniel Freedman, Danny Veikherman, Ehud Rivlin, Greg Corrado, Ilan Moshe Shimshoni, Liran Katzir, Tomer Golany, Yochai Blau, and Yossi Matias. Detecting deficient coverage in colonoscopies. *IEEE Transactions on Medical Imaging*, 2020.

Pablo Azagra, Carlos Sostres, Angel Ferrandez, Luis Riazuelo, Clara Tomasini, Oscar León Barbed, Javier Morlana, David Recasens, Victor M. Batlle, Juan J. Gómez-Rodríguez, Richard Elvira, Julia López, Cristina Oriol, Javier Civera, Juan D. Tardós, Ana Cristina Murillo, Angel Lanas, and José M. M. Montiel. Endomapper dataset of complete calibrated endoscopy procedures. 2022.

Timothy D Barfoot. *State estimation for robotics*. Cambridge University Press, 2017.

Jonathan T. Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(8):1670–1687, 2015. doi: 10.1109/TPAMI.2014.2377712.

Jonathan T. Barron, Andrew Adams, YiChang Shih, and Carlos Hernández. Fast bilateral-space stereo for synthetic defocus. *CVPR*, 2015.

Víctor M. Batlle, José M.M. Montiel, and Juan D. Tardós. Photometric single-view dense 3D reconstruction in endoscopy. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4904–4910, 2022. doi: 10.1109/IROS47612.2022.9981742.

Paul J Besl and Neil D McKay. Method for registration of 3-D shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, pages 586–606. International Society for Optics and Photonics, 1992.

Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4009–4018, 2021.

Taylor L. Bobrow, Mayank Golhar, Rohan Vijayan, Venkata Akshintala, Juan R. Garcia, and Nicholas J. Durr. Colonoscopy 3D video dataset with paired depth from 2D-3D registration. *arXiv:2206.08903*, 2022.

Mark Boss, Varun Jampani, Raphael Braun, Ce Liu, Jonathan Barron, and Hendrik PA Lensch. Neural-pil: Neural pre-integrated lighting for reflectance decomposition. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 10691–10704. Curran Associates, Inc., 2021.

Alexandre Boulch and Renaud Marlet. Deep learning for robust normal estimation in unstructured point clouds. In *Computer Graphics Forum*, volume 35, pages 281–290. Wiley Online Library, 2016.

Eric Brachmann, Martin Humenberger, Carsten Rother, and Torsten Sattler. On the limits of pseudo ground truth in visual camera re-localisation. In *CVPR*, 2021.

C. Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, José Neira, Ian Reid, and John J Leonard. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on robotics*, 32(6): 1309–1332, 2016.

Guanying Chen, Kai Han, Boxin Shi, Yasuyuki Matsushita, and Kwan-Yee K Wong. Self-calibrating deep photometric stereo networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8739–8747, 2019a.

Richard J Chen, Taylor L Bobrow, Thomas Athey, Faisal Mahmood, and Nicholas J Durr. SLAM endoscopy enhanced by adversarial depth prediction. *KDD Workshop on Applied Data Science for Healthcare*, 2019b.

Yang Chen and Gérard Medioni. Object modelling by registration of multiple range images. *Image and vision computing*, 10(3):145–155, 1992.

Yuhua Chen, Cordelia Schmid, and Cristian Sminchisescu. Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera. In *IEEE/CVF International Conference on Computer Vision*, pages 7063–7072, 2019c.

Kai Cheng, Yiting Ma, Bin Sun, Yang Li, and Xuejin Chen. Depth estimation for colonoscopy images with self-supervised learning from videos. In *Medical Image Computing and Computer Assisted Intervention–MICCAI*, 2021.

Gastone Ciuti, Marco Visentini-Scarzanella, Alessio Dore, Arianna Menciassi, Paolo Dario, and Guang-Zhong Yang. Intra-operative monocular 3D reconstruction for image-guided navigation in active locomotion capsule endoscopy. In *2012 4th IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob)*, pages 768–774, 2012. doi: 10.1109/BioRob.2012.6290771.

Toby Collins and Adrien Bartoli. Towards live monocular 3D laparoscopy using shading and specularity information. In *Int. Conf. Inf. Process. in Computer-Assisted Interventions*, pages 11–21. Springer, 2012a.

Toby Collins and Adrien Bartoli. 3D reconstruction in laparoscopy with close-range photometric stereo. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI*, pages 634–642. Springer, 2012b.

Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 303–312, 1996.

Jan Czarnowski, Tristan Laidlow, Ronald Clark, and Andrew J Davison. DeepFactors: Real-time probabilistic dense monocular SLAM. *IEEE Robotics and Automation Letters*, 5(2):721–728, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural Safety*, 31:105–112, 03 2009. doi: 10.1016/j.strusafe.2008.06.020.

Tom van Dijk and Guido de Croon. How do neural networks see depth in single images? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2183–2191, 2019.

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

H. Durrant-Whyte and T. Bailey. Simultaneous localization and mapping: part i. *IEEE Robotics Automation Magazine*, 13(2):99–110, 2006.

David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *IEEE international conference on computer vision*, 2015.

David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 2366–2374, Cambridge, MA, USA, 2014. MIT Press.

J. Engel, Thomas Schöps, and Daniel Cremers. LSD-SLAM: Large-scale direct monocular SLAM. In *European conference on computer vision*, 2014.

Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. *IEEE transactions on pattern analysis and machine intelligence*, 40(3):611–625, 2017.

Andre Esteva, Brett Kuprel, Roberto Novoa, Justin Ko, Susan Swetter, Helen Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542, 01 2017. doi: 10.1038/nature21056.

Jose M. Facil, Alejo Concha, Luis Montesano, and Javier Civera. Single-view and multi-view depth fusion. *IEEE Robotics and Automation Letters*, 2(4):1994–2001, oct 2017.

Jose M Facil, Benjamin Ummenhofer, Huizhong Zhou, Luis Montesano, Thomas Brox, and Javier Civera. CAM-Convs: camera-aware multi-scale convolutions for single-view depth. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Georgios Fagogenis, Margherita Mencattelli, Zurab Machaidze, Benoit Rosa, Karl Price, F Wu, V Weixler, Mossab Saeed, John E Mayer, and Pierre E Dupont. Autonomous robotic intracardiac catheter navigation using haptic vision. *Science robotics*, 4(29), 2019.

Rui Fan, Hengli Wang, Bohuan Xue, Huaiyang Huang, Yuan Wang, Ming Liu, and Ioannis Pitas. Three-filters-to-normal: An accurate and ultrafast surface normal estimator. *IEEE Robotics and Automation Letters*, 6(3):5405–5412, 2021.

Sebastian Farquhar, Michael Osborne, and Yarin Gal. Radial bayesian neural networks: Beyond discrete support in large-scale bayesian deep learning. *Proceedings of the 23rtd International Conference on Artificial Intelligence and Statistics*, 2020.

Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective, 2020.

Daniel Freedman, Yochai Blau, Liran Katzir, Amit Aides, Ilan Shimshoni, Danny Veikherman, Tomer Golany, Ariel Gordon, Greg Corrado, Yossi Matias, et al. Detecting deficient coverage in colonoscopies. *IEEE Transactions on Medical Imaging*, 39(11): 3451–3462, 2020.

Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2002–2011, 2018.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 2016.

Andreas Geiger, Martin Roser, and Raquel Urtasun. Efficient large-scale stereo matching. In *Asian conference on computer vision*, pages 25–38, 2010.

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE conference on computer vision and pattern recognition*, 2014.

Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3828–3838, 2019.

Juan J Gómez Rodríguez, José Lamarca, Javier Morlana, Juan D Tardós, and José MM Montiel. SD-DefSLAM: Semi-Direct Monocular SLAM for Deformable and Intracorporeal Scenes. In *IEEE International Conference on Robotics and Automation*, 2021.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Ariel Gordon, Hanhan Li, Rico Jonschkowski, and Anelia Angelova. Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In *IEEE/CVF International Conference on Computer Vision*, pages 8977–8986, 2019.

Oscar G Grasa, Javier Civera, and JMM Montiel. EKF monocular SLAM with relocalization for laparoscopic sequences. In *2011 IEEE International Conference on Robotics and Automation*, pages 4816–4821. IEEE, 2011.

Oscar G Grasa, Ernesto Bernal, Santiago Casado, Ismael Gil, and JMM Montiel. Visual SLAM for handheld monocular endoscope. *IEEE transactions on medical imaging*, 33 (1):135–146, 2013.

Xiaoyang Guo, Hongsheng Li, Shuai Yi, Jimmy Ren, and Xiaogang Wang. Learning monocular depth by distilling cross-domain stereo networks. In *European Conference on Computer Vision*, 2018.

Fredrik K Gustafsson, Martin Danelljan, and Thomas B Schon. Evaluating scalable Bayesian deep learning methods for robust computer vision. In *CVPR Workshops*, 2020.

Yang Hao, Jing Li, Fei Meng, Peisen Zhang, Gastone Ciuti, Paolo Dario, and Qiang Huang. Photometric stereo-based depth map reconstruction for monocular capsule endoscopy. *Sensors*, 20(18):5403, 2020a.

Yang Hao, Marco Visentini-Scarzanella, Jing Li, Peisen Zhang, Gastone Ciuti, Paolo Dario, and Qiang Huang. Light source position calibration method for photometric stereo in capsule endoscopy. *Advanced Robotics*, 34(12):789–801, 2020b.

Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2 edition, 2003.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Berthold K.P. Horn and Michael J. Brooks, editors. *Shape from Shading*. MIT Press, 1989.

Baoru Huang, Jian-Qing Zheng, Anh Nguyen, David Tuch, Kunal Vyas, Stamatia Giannarou, and Daniel S Elson. Self-supervised generative adversarial network for depth estimation in laparoscopic images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI*, 2021.

Seung-Jun Hwang, Sung-Jun Park, Gyu-Min Kim, and Joong-Hwan Baek. Unsupervised monocular depth estimation for colonoscope system using feedback network. *Sensors*, 21 (8), 2021.

Eddy Ilg, Ozgun Cicek, Silvio Galesso, Aaron Klein, Osama Makansi, Frank Hutter, and Thomas Brox. Uncertainty estimates and multi-hypotheses networks for optical flow. In *ECCV*, 2018.

Kağan İncetan, Ibrahim Omer Celik, Abdulhamid Obeid, Guliz Irem Gokceler, Kutsev Bengisu Ozyoruk, Yasin Almalioglu, Richard J Chen, Faisal Mahmood, Hunter Gilbert, Nicholas J Durr, et al. VR-Caps]: a virtual environment for capsule endoscopy. *Medical image analysis*, 70:101990, 2021.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

Yuji Iwahori, Shun Emoto, Kenji Funahashi, M. K. Bhuyan, Aili Wang, and Kunio Kasugai. Recovering shape and size from a single endoscope image using optimization. In *2022 12th International Congress on Advanced Applied Informatics (IIAI-AAI)*, pages 331–334, 2022. doi: 10.1109/IIAIAAI55812.2022.00073.

Sergio Izquierdo and Javier Civera. Sfm-ttr: Using structure from motion for test-time refinement of single-view depth networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

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

Rawen Kader, Anton Cid-Mejias, Patrick Brandao, Shahraz Islam, Sanjith Hebbar, Juana González-Bueno Puyal, Omer F. Ahmad, Mohamed Hussein, Daniel Toth, Peter Mountney, Ed Seward, Roser Vega, Danail Stoyanov, and Laurence B. Lovat. Polyp characterization using deep learning and a publicly accessible polyp video database. *Digestive Endoscopy*, 35(5):645–655, 2023. doi: https://doi.org/10.1111/den.14500.

Mert Asim Karaoglu, Nikolas Brasch, Marijn Stollenga, Wolfgang Wein, Nassir Navab, Federico Tombari, and Alexander Ladikos. Adversarial domain feature adaptation for bronchoscopic depth estimation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI*, pages 300–310. Springer, 2021.

Elia Kaufmann, Leonard Bauersfeld, Antonio Loquercio, Matthias Müller, Vladlen Koltun, and Davide Scaramuzza. Champion-level drone racing using deep reinforcement learning. *Nature*, 620(7976):982–987, 2023.

Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? *NeurIPS*, 2017.

Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *CoRR*, abs/1511.02680, 2015a.

Alex Kendall, Matthew Grimes, and Roberto Cipolla. PoseNet: A convolutional network for real-time 6-DOF camera relocalization. In *Proceedings of the IEEE international conference on computer vision*, pages 2938–2946, 2015b.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491, 2018.

Mohammad Emtiyaz Khan, Didrik Nielsen, Voot Tangkaratt, Wu Lin, Yarin Gal, and Akash Srivastava. Fast and scalable Bayesian deep learning by weight-perturbation in Adam. *35th International Conference on Machine Learning, ICML 2018*, 6:4088–4113, 2018.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. 12 2014.

Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural Safety*, 31(2):105–112, 2009. ISSN 0167-4730. Risk Acceptance and Risk Communication.

Maria Klodt and Andrea Vedaldi. Supervising the new with the old: learning sfm from sfm. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

Yevhen Kuznietsov, Jorg Stuckler, and Bastian Leibe. Semi-supervised deep learning for monocular depth map prediction. In *IEEE conference on computer vision and pattern recognition*, 2017.

Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*, pages 239–248, 2016.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, pages 6403–6414, 2017. ISSN 10495258.

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, 2021. doi: 10.1109/TRO.2020.3020739.

Jose Lamarca and Jose Maria Martinez Montiel. Camera tracking for SLAM in deformable maps. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018.

Katrin Lasinger, René Ranftl, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *CoRR*, abs/1907.01341, 2019.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553): 436–444, 2015.

Jae-Han Lee, Minhyeok Heo, Kyung-Rae Kim, and Chang-Su Kim. Single-image depth estimation based on fourier domain analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 330–339, 2018. doi: 10.1109/CVPR.2018.00042.

Seong Hun Lee and Javier Civera. Loosely-coupled semi-direct monocular SLAM. *IEEE Robotics and Automation Letters*, 4(2):399–406, 2018.

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

Zhaoshuo Li, Nathan Drenkow, Hao Ding, Andy S Ding, Alexander Lu, Francis X Creighton, Russell H Taylor, and Mathias Unberath. On the sins of image synthesis loss for self-supervised depth estimation. *arXiv preprint arXiv:2109.06163*, 2021.

Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2041–2050, 2018.

Zhengqin Li, Zexiang Xu, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Learning to reconstruct shape and spatially-varying reflectance from a single image. *ACM Trans. Graph.*, 37(6), dec 2018. ISSN 0730-0301. doi: 10.1145/3272127.3275055.

Zhengqin Li, Mohammad Shafiei, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and svbrdf from a single image. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2475–2484, 2020.

Zhenyu Li, Xuyang Wang, Xianming Liu, and Junjun Jiang. BinsFormer: Revisiting adaptive bins for monocular depth estimation. *arXiv preprint arXiv:2204.00987*, 2022.

Daniel Lichy, Jiaye Wu, Roni Sengupta, and David W Jacobs. Shape and material capture at home. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6123–6133, 2021.

Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian 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.

Xingtong Liu, Ayushi Sinha, Mathias Unberath, Masaru Ishii, Gregory D Hager, Russell H Taylor, and Austin Reiter. Self-supervised learning for dense depth estimation in monocular endoscopy. In *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*. 2018.

Xingtong Liu, Ayushi Sinha, Masaru Ishii, Gregory D. Hager, Austin Reiter, Russell H. Taylor, and Mathias Unberath. Dense depth estimation in monocular endoscopy with self-supervised learning methods. *IEEE Transactions on Medical Imaging*, 39(5): 1438–1447, 2020. doi: 10.1109/TMI.2019.2950936.

William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3D surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987.

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

Wenjie Luo, Alexander G Schwing, and Raquel Urtasun. Efficient deep learning for stereo matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5695–5703, 2016.

Xuan Luo, Jia-Bin Huang, Richard Szeliski, Kevin Matzen, and Johannes Kopf. Consistent video depth estimation. *ACM Transactions on Graphics (TOG)*, 39(4):71–1, 2020.

R. Ma, Rui Wang, Stephen Pizer, Julian Rosenman, Sarah K McGill, and Jan-Michael Frahm. Real-time 3D reconstruction of colonoscopic surfaces for determining missing regions. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2019.

Ruibin Ma, Rui Wang, Yubo Zhang, Stephen Pizer, Sarah K McGill, Julian Rosenman, and Jan-Michael Frahm. RNNSLAM: Reconstructing the 3D colon to visualize missing regions during a colonoscopy. *Medical image analysis*, 72:102100, 2021.

Reza Mahjourian, Martin Wicke, and Anelia Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. 2 2018.

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

Faisal Mahmood and Nicholas J Durr. Deep learning and conditional random fields-based depth estimation and topographical reconstruction from conventional endoscopy. *Medical image analysis*, 48:230–243, 2018.

Nader Mahmoud, Óscar G Grasa, Stéphane A Nicolau, Christophe Doignon, Luc Soler, Jacques Marescaux, and JMM Montiel. On-patient see-through augmented reality based on visual SLAM. *International journal of computer assisted radiology and surgery*, 12 (1):1–11, 2017.

Nader Mahmoud, Toby Collins, Alexandre Hostettler, Luc Soler, Christophe Doignon, and Jose Maria Martinez Montiel. Live tracking and dense reconstruction for handheld monocular endoscopy. *IEEE transactions on medical imaging*, 38(1):79–89, 2018.

A. Marmol, Artur Banach, and Thierry Peynot. Dense-ArthroSLAM: Dense intra-articular 3-D reconstruction with robust localization prior for arthroscopy. *IEEE Robotics and Automation Letters*, 4(2):918–925, 2019.

Shawn Mathew, Saad Nadeem, Sruti Kumari, and Arie Kaufman. Augmenting colonoscopy using extended and directional CycleGAN for lossy image translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4696–4705, June 2020.

Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4040–4048, 2016.

John McCormac, Ankur Handa, Stefan Leutenegger, and Andrew J Davison. Scenenet rgb-d: 5m photorealistic images of synthetic indoor trajectories with ground truth. *arXiv preprint arXiv:1612.05079*, 2016.

S. Mahdi H. Miangoleh, Sebastian Dille, Long Mai, Sylvain Paris, and Yağız Aksoy. Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging. 2021.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

Richard Modrzejewski, Toby Collins, Alexandre Hostettler, Jacques Marescaux, and Adrien Bartoli. Light modelling and calibration in laparoscopy. *Int. J. Computer Assisted Radiology and Surgery*, 15(5):859–866, 2020.

Peter Mountney, Danail Stoyanov, and Guang-Zhong Yang. Three-dimensional tissue deformation recovery and tracking. *IEEE Signal Processing Magazine*, 27(4):14–24, 2010.

Jishnu Mukhoti and Yarin Gal. Evaluating Bayesian deep learning methods for semantic segmentation. *arXiv preprint arXiv:1811.12709*, 2018.

Raul Mur-Artal and Juan 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.

Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

Richard A Newcombe, Steven J Lovegrove, and Andrew J Davison. DTAM: Dense tracking and mapping in real-time. In *2011 International Conference on Computer Vision*, 2011.

Richard A Newcombe, Dieter Fox, and Steven M Seitz. DynamicFusion: Reconstruction and tracking of non-rigid scenes in real-time. In *IEEE conference on computer vision and pattern recognition*, 2015.

Takayuki Okatani and Koichiro 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.

Olga Russakovsky et al. ImageNet Large Scale Visual Recognition Challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Kazuki Osawa, Siddharth Swaroop, Mohammad Emtiyaz E Khan, Anirudh Jain, Runa Eschenhagen, Richard E Turner, and Rio Yokota. Practical deep learning with Bayesian principles. *NeurIPS*, 2019.

Kutsev Bengisu Ozyoruk, Guliz Irem Gokceler, Taylor L Bobrow, Gulfize Coskun, Kagan Incetan, Yasin Almalioglu, Faisal Mahmood, Eva Curto, Luis Perdigoto, Marina Oliveira, et al. EndoSLAM dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos. *Medical Image Analysis*, 71:102058, 2021.

S. Parashar, D. Pizarro, and A. Bartoli. Isometric Non-Rigid Shape-from-Motion with Riemannian Geometry Solved in Linear Time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(10):2442–2454, 2018. doi: 10.1109/TPAMI.2017.2760301.

Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Colored point cloud registration revisited. In *IEEE international conference on computer vision*, 2017.

Vicente Parot, Daryl Lim, German Gonzalez, Giovanni Traverso, Norman S. Nishioka, Benjamin J. Vakoc, and Nicholas J. Durr. Photometric stereo endoscopy. *Journal of Biomedical Optics*, 18(7):076017, 2013.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

Vaishakh Patil, Christos Sakaridis, Alexander Liniger, and Luc Van Gool. P3Depth: Monocular depth estimation with a piecewise planarity prior. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1610–1621, June 2022.

Matteo Poggi, Filippo Aleotti, Fabio Tosi, and Stefano Mattoccia. On the uncertainty of self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3227–3237, 2020.

Janis Postels, Francesco Ferroni, Huseyin Coskun, Nassir Navab, and Federico Tombari. Sampling-free epistemic uncertainty estimation using approximated variance propagation. *CoRR*, abs/1908.00598, 2019.

Emmanuel Prados and Olivier Faugeras. Shape from shading: a well-posed problem? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 870–877, 2005. doi: 10.1109/CVPR.2005.319.

Philip Pratt, Danail Stoyanov, Marco Visentini-Scarzanella, and Guang-Zhong Yang. Dynamic guidance for robotic surgery using image-constrained biomechanical models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 77–85, 2010.

Tong Qin, Peiliang Li, and Shaojie Shen. Vins-mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Transactions on Robotics*, 34(4):1004–1020, 2018.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12179–12188, October 2021.

Anita Rau, PJ Eddie Edwards, Omer F Ahmad, Paul Riordan, Mirek Janatka, Laurence B Lovat, and Danail Stoyanov. Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy. *International Journal of Computer Assisted Radiology and Surgery*, pages 1–10, 2019.

Anita Rau, Binod Bhattarai, Lourdes Agapito, and Danail Stoyanov. Bimodal camera pose prediction for endoscopy. *arXiv preprint arXiv:2204.04968*, 2022.

David Recasens, José Lamarca, José M Fácil, JMM Montiel, and Javier Civera. Endo-depth-and-motion: Reconstruction and tracking in endoscopic videos using depth networks and photometric constraints. *IEEE Robotics and Automation Letters*, 6(4), 2021.

J. Van Rijn, Johannes B Reitsma, Jaap Stoker, Patrick M Bossuyt, Sander J Van Deventer, and Evelien Dekker. Polyp miss rate determined by tandem colonoscopy: a systematic review. *Official journal of the American College of Gastroenterology*, 101(2):343–350, 2006.

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

Javier Rodriguez-Puigvert, David Recasens, Javier Civera, and Ruben Martinez-Cantin. On the uncertain single-view depths in colonoscopies. In *Medical Image Computing and Computer Assisted Intervention – MICCAI*, pages 130–140, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-16437-8.

Javier Rodríguez-Puigvert, Víctor M. Batlle, José María M. Montiel, Rubén Martínez-Cantín, Pascal Fua, Juan D. Tardós, and Javier Civera. LightDepth: single-view depth self-supervision from illumination decline. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Shen Sang and Manmohan Chandraker. Single-shot neural relighting and svbrdf estimation. In *European Conference Computer Vision (ECCV)*, pages 85–101. Springer, 2020.

Ashutosh Saxena, Sung Chung, and Andrew Ng. Learning depth from single monocular images. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems*, volume 18. MIT Press, 2005.

Ashutosh Saxena, Min Sun, and Andrew Y Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE transactions on pattern analysis and machine intelligence*, 31 (5):824–840, 2008.

Davide Scaramuzza, Agostino Martinelli, and Roland Siegwart. A toolbox for easily calibrating omnidirectional cameras. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5695–5701, 2006. doi: 10.1109/IROS.2006.282372.

Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016.

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

Agniva Sengupta and Adrien Bartoli. Colonoscopic 3D reconstruction by tubular non-rigid structure-from-motion. *International Journal of Computer Assisted Radiology and Surgery*, 16(7):1237–1241, Jul 2021. ISSN 1861-6429. doi: 10.1007/s11548-021-02409-x.

Roni Sengupta, Angjoo Kanazawa, Carlos D. Castillo, and David W. Jacobs. Sfsnet: Learning shape, refectance and illuminance of faces in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Regognition (CVPR)*, 2018.

Lalith Sharan, Lukas Burger, Georgii Kostiuchik, Ivo Wolf, Matthias Karck, Raffaele De Simone, and Sandy Engelhardt. Domain gap in adapting self-supervised depth estimation methods for stereo-endoscopy. *Current Directions in Biomedical Engineering*, 6(1), 2020.

Mali Shen, Yun Gu, Ning Liu, and Guang-Zhong Yang. Context-aware depth and pose estimation for bronchoscopic navigation. *IEEE Robotics and Automation Letters*, 4(2), 2019.

Chang Shu, Kun Yu, Zhixiang Duan, and Kuiyuan Yang. Feature-metric loss for self-supervised learning of depth and egomotion. In *European Conference Computer Vision (ECCV)*, pages 572–588. Springer, 2020.

J. Song, Jun Wang, Liang Zhao, Shoudong Huang, and Gamini Dissanayake. Dynamic reconstruction of deformable soft-tissue with stereo scope in minimal invasive surgery. *IEEE Robotics and Automation Letters*, 3(1):155–162, 2017.

Jingwei Song, Jun Wang, Liang Zhao, Shoudong Huang, and Gamini Dissanayake. MIS-SLAM: Real-time large-scale dense deformable SLAM system in minimal invasive surgery based on heterogeneous computing. *IEEE Robotics and Automation Letters*, 3(4): 4068–4075, 2018.

M. Song, S. Lim, and W. Kim. Monocular depth estimation using laplacian pyramid-based depth residuals. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(11): 4381–4393, 2021. doi: 10.1109/TCSVT.2021.3049869.

Pratul P. Srinivasan, Tongzhou Wang, Ashwin Sreelal, Ravi Ramamoorthi, and Ren Ng. Learning to synthesize a 4d rgbd light field from a single image. In *IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

Frank Steinbrücker, Jürgen Sturm, and Daniel Cremers. Real-time visual odometry from dense RGB-D images. In *2011 IEEE international conference on computer vision workshops (ICCV Workshops)*, 2011.

Danail Stoyanov, George P Mylonas, Fani Deligianni, Ara Darzi, and Guang Zhong Yang. Soft-tissue motion tracking and structure estimation for robotic assisted mis procedures. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 139–146, 2005.

Danail Stoyanov, Marco Visentini Scarzanella, Philip Pratt, and Guang-Zhong Yang. Real-time stereo reconstruction in robotically assisted minimally invasive surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2010.

V. B. Surya Prasath and Hiroharu Kawanaka. Near-light perspective shape from shading for 3D visualizations in endoscopy systems. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2293–2295, 2017. doi: 10.1109/BIBM.2017.8218031.

Keisuke Tateno, Federico Tombari, Iro Laina, and Nassir Navab. CNN-SLAM: Real-time dense monocular SLAM with learned depth prediction. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Ayush Tewari, Justus Thies, Ben Mildenhall, Pratul Srinivasan, Edgar Tretschk, Wang Yifan, Christoph Lassner, Vincent Sitzmann, Ricardo Martin-Brualla, Stephen Lombardi, et al. Advances in neural rendering. In *Computer Graphics Forum*, volume 41, pages 703–735. Wiley Online Library, 2022.

Lokender Tiwari, Pan Ji, Quoc-Huy Tran, Bingbing Zhuang, Saket Anand, and Manmohan Chandraker. Pseudo RGB-D for self-improving monocular SLAM and depth prediction. In *European Conference on Computer Vision*, pages 437–455. Springer, 2020.

Seiya Tsuda, Yuji Iwahori, M. K. Bhuyan, Robert J. Woodham, and Kunio Kasugai. Recovering 3D shape with absolute size from endoscope images using RBF neural network. *Journal of Biomedical Imaging*, 2015, jan 2015. ISSN 1687-4188. doi: 10.1155/2015/109804.

Mehmet Turan, Evin Pinar Ornek, Nail Ibrahimli, Can Giracoglu, Yasin Almalioglu, Mehmet Fatih Yanik, and Metin Sitti. Unsupervised odometry and depth learning for endoscopic capsule robots. In *IROS*, 2018.

Benjamin Ummenhofer, Jonas Uhrig, Nikolaus Mayer, and Eddy Ilg. Demon: Depth and motion network for learning monocular stereo.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

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

Marco Visentini-Scarzanella, Danail Stoyanov, and Guang-Zhong Yang. Metric depth recovery from monocular images using shape-from-shading and specularities. In *IEEE International Conference on Image Processing*, pages 25–28, 2012. doi: 10.1109/ICIP.2012.6466786.

Marco Visentini-Scarzanella, Takamasa Sugiura, Toshimitsu Kaneko, and Shinichiro Koto. Deep monocular 3D reconstruction for assisted navigation in bronchoscopy. *International Journal of Computer Assisted Radiology and Surgery*, 12(7):1089–1099, Jul 2017. ISSN 1861-6429. doi: 10.1007/s11548-017-1609-2.

Ning-Hsu Wang, Ren Wang, Yu-Lun Liu, Yu-Hao Huang, Yu-Lin Chang, Chia-Ping Chen, and Kevin Jou. Bridging unsupervised and supervised depth from focus via all-in-focus supervision. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12621–12631, October 2021.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Jamie Watson, Oisin Mac Aodha, Victor Prisacariu, Gabriel Brostow, and Michael Firman. The temporal opportunist: Self-supervised multi-frame monocular depth. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1164–1174, 2021.

T. Whelan, Renato F Salas-Moreno, Ben Glocker, Andrew J Davison, and Stefan Leutenegger. ElasticFusion: Real-time dense SLAM and light source estimation. *The International Journal of Robotics Research*, 35(14):1697–1716, 2016.

Aji Resindra Widya, Yusuke Monno, Masatoshi Okutomi, Sho Suzuki, Takuji Gotoda, and Kenji Miki. Self-supervised monocular depth estimation in gastroendoscopy using gan-augmented images. In *Medical Imaging 2021: Image Processing*, 2021.

Andrew Gordon Wilson. The case for bayesian deep learning. *arXiv preprint arXiv:2001.10995*, 2020.

Chenyu Wu, Srinivasa G. Narasimhan, and Branislav Jaramaz. A multi-image shape-from-shading framework for near-lighting perspective endoscopes. *International Journal of Computer Vision*, 86(2):211–228, Jan 2010. ISSN 1573-1405. doi: 10.1007/s11263-009-0207-3.

Ke Xian, Jianming Zhang, Oliver Wang, Long Mai, Zhe Lin, and Zhiguo Cao. Structure-guided ranking loss for single image depth prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

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

Jiaxing Yan, Hong Zhao, Penghui Bu, and YuSheng Jin. Channel-wise attention-based network for self-supervised monocular depth estimation. In *IEEE International Conference on 3D vision (3DV)*, pages 464–473, 2021.

Qingan Yan, Pan Ji, Nitin Bansal, Yuxin Ma, Yuan Tian, and Yi Xu. Fisheyedistill: Self-supervised monocular depth estimation with ordinal distillation for fisheye cameras. *arXiv preprint arXiv:2205.02930*, 2022.

Gengshan Yang, Peiyun Hu, and Deva Ramanan. Inferring distributions over depth from a single image. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6090–6096. IEEE, 2019.

Zhenheng Yang, Peng Wang, Wei Xu, Liang Zhao, and Ramakant 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.

Huangying Zhan, Ravi Garg, Chamara Saroj Weerasekera, Kejie Li, Harsh Agarwal, and Ian Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *CVPR*, 2018.

Ruo Zhang, Ping-Sing Tsai, J.E. Cryer, and M. Shah. Shape-from-shading: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(8):690–706, 1999. doi: 10.1109/34.784284.

Shuai Zhang, Liang Zhao, Shoudong Huang, Menglong Ye, and Qi 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.

Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM Transactions on Graphics (TOG)*, 40(6):1–18, 2021.

Yuanqing Zhang, Jiaming Sun, Xingyi He, Huan Fu, Rongfei Jia, and Xiaowei Zhou. Modeling indirect illumination for inverse rendering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18643–18652, 2022.

Huizhong Zhou, Benjamin Ummenhofer, and Thomas Brox. Deeptam: Deep tracking and mapping. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11220 LNCS:851–868, 2018a. ISSN 16113349.

Huizhong Zhou, Benjamin Ummenhofer, and Thomas Brox. DeepTAM: Deep tracking and mapping with convolutional neural networks. *International Journal of Computer Vision*, 128(3):756–769, 2020a.

Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3D: A modern library for 3D data processing. *arXiv:1801.09847*, 2018b.

Qunjie Zhou, Torsten Sattler, Marc Pollefeys, and Laura Leal-Taixe. To learn or not to learn: Visual localization from essential matrices. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3319–3326. IEEE, 2020b.

Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1851–1858, 2017.

Zhongkai Zhou, Xinnan Fan, Pengfei Shi, and Yuanxue Xin. R-msfm: Recurrent multi-scale feature modulation for monocular depth estimating. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12777–12786, October 2021.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

# List of Figures

# List of Tables