

POLYTECHNIC OF TURIN

Master's Degree in Biomedical Engineering



**Politecnico
di Torino**

Master's Degree Thesis

3D Reconstruction of the Colonic Mucosa from Monocular Endoscopic Video for Unobserved Area Quantification

Supervisors

Prof. Kristen M. MEIBURGER

Prof. Alberto AREZZO

Eng. Francesco MARZOLA

Candidate

LORENZO REVELLO

December 2025

Declaration

I hereby declare that the content and organization of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

This thesis is licensed under a Creative Commons License, Attribution - Noncommercial - NoDerivative works 4.0 International: see www.creativecommons.org. The text may be reproduced for non-commercial purposes, provided that credit is given to the original author.

Turin, 4/12/2025

AI Use Disclosure

In the writing process of this thesis the following tools have been used to improve the readability and language:

- ChatGPT (OpenAI)
- Google AI Studio (Google)

Turin, 4/12/2025

Abstract

Colorectal cancer remains one of the leading causes of cancer mortality. Colonoscopy is the clinical gold standard for detecting and removing precancerous polyps because it provides direct visualization of the colonic mucosa. Nevertheless, the field of view is limited and the colon’s geometry is complex, so relevant areas can remain unobserved during routine inspection.

This thesis introduces a pipeline that reconstructs a three-dimensional model of the inner colonic surface from monocular endoscopic video, with the aim of measuring where and how much mucosa was not seen.

The method fuses RGB frames, predicted depth, and estimated camera pose through a TSDF-based reconstruction module to produce dense, anatomically faithful meshes. Depth is inferred using foundation predictors (Depth Anything, Video Depth Anything, DepthPro) and a SUMNet model trained on the SimCol3D dataset; poses are estimated with a bimodal deep learning-based model trained on SimCol3D data. Reconstructions were generated from 15 virtual colonoscopies under different configurations, and an ablation study compared meshes built from estimated inputs against a GT-depth + GT-pose baseline. Three cases were evaluated: (i) estimated depth + GT pose, (ii) GT depth + estimated pose, and (iii) both estimated. Reconstructed meshes were aligned to the GT baseline, and performance was assessed using the symmetric Chamfer-L1 distance and mesh-to-mesh overlap at tolerance at $\tau = 4$ mm.

- (i) C-L1 = 4.50 mm, mesh-to-mesh overlap @ $\tau = 82.3\%$.
- (ii) C-L1 = 38.13 mm, mesh-to-mesh overlap @ $\tau = 35.7\%$.
- (iii) C-L1 = 41.18 mm, mesh-to-mesh overlap @ $\tau = 27.9\%$.

Missing regions are identified by closing reconstructed meshes with the Poisson Surface Reconstruction method and comparing the closed mesh to the original one. The detected missing regions are analyzed for their extent and size. Across all acquisitions, the percentage of unobserved surface averages $19.6\% \pm 1.9\%$, with unobserved areas clustering around deep folds and sharp flexures.

To bridge simulation and reality, a hybrid dataset is acquired on a silicone colon phantom. An Olympus EVIS EXERA III endoscope captures fisheye video sequences, while an NDI AURORA electromagnetic tracker records six-degree-of-freedom trajectories of the endoscope tip within the operative field. RGB frames are paired with depth maps estimated using the Depth-Anything model, providing a multimodal dataset linking visual and spatial data.

Some limitations remain: foundation depth models underperform on endoscopic

video due to domain gaps (fluids, specular mucosa, low texture) and monocular scale ambiguity; colon deformability is treated as quasi-rigid; and the pipeline is not yet real-time. Despite these constraints, the framework produces smooth, anatomically coherent reconstructions, quantifies unseen mucosa with a simple metric, and shows clear gains from domain-specific fine-tuning. These results establish a practical basis for spatially aware analysis of colonoscopy, support objective coverage reporting, and lay the groundwork for future real-time clinical deployment.

Acknowledgements

I would like to begin by expressing my thanks to those who made this thesis possible, contributing decisively to my professional and personal growth.

My first thanks go to my supervisor, Prof. Meiburger. I thank her for the trust she placed in my work, and for her availability and professionalism throughout the entire process.

I would like to mention the Polytechnic of Turin, the institution that has trained me. My university experience has been an intense and stimulating journey, providing me with the foundations and tools to face future challenges with competence and curiosity.

A special dedication goes to the MITIC laboratory, which welcomed me and has been an important environment for my growth.

In this context, a particular thank you goes to Prof. Arezzo. I thank him not only for the opportunity to develop my thesis in a state-of-the-art research center, but above all for being a source of inspiration. His vision, passion, and ability to motivate are contagious and allowed me to deeply appreciate the value of our work. Thanks to him, I had the privilege of coming into direct contact with the hospital environment, working in a unique interdisciplinary setting surrounded by surgeons and engineers.

A fundamental thank you is dedicated to Eng. Marzola, who closely guided me in every technical aspect of this project. Thank you for your patience during difficult times, for the continuous exchange of ideas, and for your constant availability. You have been an indispensable point of reference.

Table of Contents

List of Tables	x
List of Figures	xI
1 Introduction	1
1.1 Context	1
1.1.1 Colorectal cancer and precursor polyps	1
1.1.2 Colonoscopy	2
1.1.3 Colon anatomy	3
1.1.4 Monocular endoscopy	4
1.2 Motivation of the thesis	5
1.3 Aim of the thesis	6
1.4 Contribution of the thesis	7
1.4.1 Development of a 3D reconstruction system	7
1.4.2 Implementation of a mesh analysis framework for missing- region detection	7
1.4.3 Comparative study of depth estimation models	7
1.4.4 Acquisition of a real dataset combining video frames and camera poses	8
2 Background	9
2.1 3D reconstruction in medicine: overview and applications	9
2.1.1 Traditional techniques	10
2.1.2 Video-based techniques	11
2.2 3D reconstruction techniques	12
2.2.1 Geometric multi-view methods	12
2.2.2 Photometric methods	15
2.2.3 Learning based methods	18
2.3 3D Representation and Visualization	21
2.3.1 Point clouds	21
2.3.2 Polygonal meshes	22

2.3.3	Truncated Signed Distance Fields (TSDF)	23
2.3.4	Neural Radiance Fields (NeRF)	24
2.3.5	3D Gaussian Splatting (3DGS)	26
2.4	Missing regions identification	28
2.4.1	Coverage in CT colonography	28
2.4.2	Reconstruction-based coverage	28
2.4.3	Real-time guidance toward unseen mucosa	28
2.4.4	Quantifying coverage on reconstructed surfaces	28
2.5	Related works for colon 3D reconstruction	29
2.5.1	RNN-SLAM	29
2.5.2	ColVO	29
2.5.3	Endo2DTAM	31
2.5.4	EndoGSLAM	32
2.5.5	C ³ Fusion	32
2.5.6	Summary	34
3	Materials and Methods	35
3.1	Datasets	35
3.1.1	SimCol3D	35
3.1.2	C3VD	36
3.2	TSDF-module	37
3.2.1	Inputs	37
3.2.2	Data processing	39
3.2.3	Pipeline	40
3.2.4	Outputs	41
3.3	Missing regions analysis	43
3.3.1	Inputs	43
3.3.2	Mesh closing	43
3.3.3	Missing region identification and quantification	44
3.3.4	Missing regions distribution	49
3.4	Depth estimation	51
3.4.1	Depth-Pro	51
3.4.2	Depth-Anything	52
3.4.3	Video-Depth-Anything	53
3.4.4	SUMNet	53
3.4.5	Depth estimation summary	55
3.4.6	Depth processing	56
3.5	Pose estimation	57
3.6	Hybrid Dataset acquisition	60
3.6.1	Endoscope and video acquisition	60
3.6.2	Phantom model	60

3.6.3	Aurora NDI	61
3.6.4	Pose computation	62
3.6.5	Endoscope camera calibration	64
4	Results	65
4.1	TSDF validation	65
4.2	Depth prediction	66
4.3	Pose prediction	69
4.3.1	Quantitative pose estimation results	69
4.4	Missing regions	73
4.4.1	Percentage of unobserved surface	73
4.4.2	Distribution of missing regions	74
4.4.3	Uncertainty map of the unobserved regions	76
4.5	Ablation study - Mesh to Mesh	78
4.6	Aurora Dataset	83
4.6.1	Dataset	84
5	Discussion	86
5.1	Interpretation of Results	86
5.2	Clinical Implications and Potential Applications	91
5.3	Methodological Reflections and Design Choices	91
5.4	Limitations	92
5.5	Future Work	94
6	Conclusions	96
6.1	Summary	96
6.2	Concluding Remarks	97
	Bibliography	99

List of Tables

2.1	Comparison of 3D reconstruction methods of the colonic surface. . .	34
4.1	Depth estimation metrics on SimCol3D SyntheticColon_I (15 sequences). MAE and RMSE are both expressed in millimeters	68
4.2	Pose estimation results aggregated by data split.	72
4.3	Unobserved regions for SyntheticColonI.	73
4.4	Unobserved regions for SyntheticColonII.	74
4.5	Statistics of the 20 largest unobserved surface regions.	76
4.6	Mesh-to-mesh comparison metrics for ablation study (averaged over 6 test sequences).	81
4.7	Acquired dataset.	83
4.8	Camera intrinsic parameters.	84

List of Figures

1.1	Colonoscopy procedure	2
1.2	Colon anatomy	3
1.3	Monocular endoscope setup	4
2.1	Structure from Motion principle	13
2.2	SLAM principle	15
2.3	Shape from Shading principle	16
2.4	Photometric Stereo principle	17
2.5	NeRF principle	25
2.6	Gaussian Splatting principle	27
2.7	Pipeline of RNN-SLAM	29
2.8	3D reconstruction results of RNN-SLAM	29
2.9	Pipeline of ColVO	30
2.10	3D reconstruction results of ColVO	30
2.11	Pipeline of Endo2DTAM	31
2.12	3D reconstruction results of Endo2DTAM	31
2.13	Pipeline of EndoGSLAM	32
2.14	3D reconstruction results of EndoGSLAM	32
2.15	Pipeline of C ³ Fusion	33
2.16	3D reconstruction results of C ³ Fusion	33
3.1	3D models of the three synthetic colons the SimCol3D dataset. (SyntheticColon_I, SyntheticColon_II, SyntheticColon_III)	36
3.2	C3VD dataset	37
3.3	TSDF 3D reconstruction pipeline	42
3.4	Missing region identification pipeline.	47
3.5	Missing regions centroid and trajectory nearest point	49
3.6	Heatmap pipeline	50
3.7	Depth-Pro pipeline	51
3.8	Depth-Anything pipeline	52
3.9	Video-Depth-Anything pipeline	53

3.10	SUMNet architecture	54
3.11	Bimodal camera pose prediction architecture (Class Net + Pose Net with correlation volume and residuals around class means).	58
3.12	Aurora components. From left to right: FG, SCU, SIU.	62
3.13	Aurora acquisition experimental setup	63
3.14	5 of the 20 endoscope images used for calibration.	64
3.15	Detected checkerboard corners in the calibration images.	64
4.1	3D reconstruction obtained from a C3VD video (colored) compared with the GT model (gray)	65
4.2	Examples of depth prediction on <i>SimCol3D</i> sequences.	68
4.3	Predicted camera trajectories on SimCol3D sequences	69
4.4	Heatmap of the missing regions across 15 sequences, the zoomed regions highlight the areas most frequently unobserved	77
4.5	Mesh-to-mesh comparison for ablation study	82
4.6	Sample frames from Aurora dataset.	85

Chapter 1

Introduction

1.1 Context

1.1.1 Colorectal cancer and precursor polyps

Colorectal cancer (CRC) is one of the most common and deadliest cancers worldwide. In 2022 alone, the global burden included over 1.9 million new CRC cases and sadly, more than 900,000 deaths globally.¹ The severity of CRC is often compounded by the fact that the disease is frequently asymptomatic in its early stages. This lack of clear symptoms often leads to delayed diagnosis, which significantly compromises the patient's prognosis and limits the available therapeutic options.

The vast majority of CRCs originate from polyps, which are small growths or masses that form on the inner lining of the colon or rectum. While polyps are non-cancerous themselves, they are classified as pre-cancerous lesions, meaning they possess the potential to evolve into invasive cancer over a period of time if left untreated.

During established screening or diagnostic examinations, such as colonoscopy, the primary goal is the timely detection and identification of these polyps. Once located, polyps are routinely removed via minimally invasive procedures (polypectomy). The excised tissue is then immediately sent for pathological examination, which is crucial for determining its histological features, assessing the degree of dysplasia (abnormal cell growth), and establishing the effective risk of future malignancy. The prompt and complete removal of these pre-cancerous polyps is, therefore, the most successful strategy to drastically reduce the likelihood of subsequent cancer development.

¹<https://www.iarc.who.int/cancer-type/colorectal-cancer>

1.1.2 Colonoscopy

Colonoscopy is the gold standard for both finding and therapeutically removing colonic polyps, typically achieved within a single clinical session. The procedure involves the insertion of a flexible colonoscope, navigating the large intestine from the rectum to the cecum. The instrument is equipped with a high-resolution camera that transmits a live video feed to a monitor, allowing the endoscopist to perform a detailed, real-time inspection of the mucosal lining.

The colonoscope features an instrument channel through which specialized tools can be passed. If the endoscopist identifies a suspicious area or polyp, that lesion can be immediately acted upon. Small lesions can be biopsied for tissue diagnosis, while pre-cancerous polyps are removed immediately using techniques like snare polypectomy. In practice, the exam includes three critical stages: thorough bowel preparation, the insertion phase to reach the cecum, and most importantly, a slow and meticulous withdrawal phase focused on detailed mucosal inspection to ensure no lesion is missed.

The primary advantages of colonoscopy are manifold: direct visualization of the entire bowel anatomy, immediate therapeutic intervention, precise tissue acquisition, and the ability to document findings with high-quality images or video clips. These features make colonoscopy central to prevention and early detection strategies, allowing clinicians to interrupt the malignant polyp-to-cancer pathway at the moment of discovery.

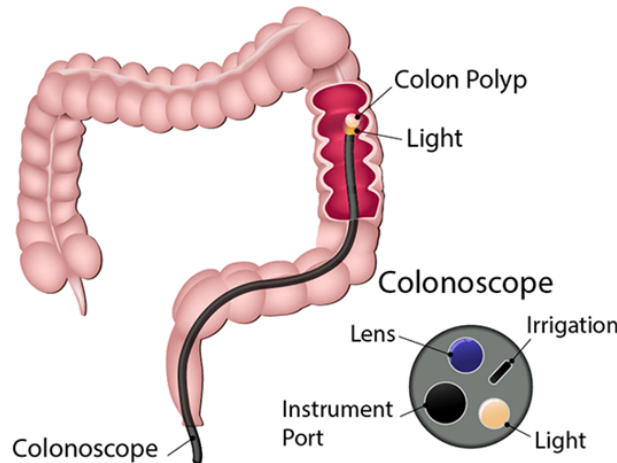


Figure 1.1: Colonoscopy procedure²

²<https://www.ibdrelief.com/learn/diagnosis/tests/endoscopy-tests-for-ibd/colonoscopy-for-ibd>

1.1.3 Colon anatomy

The colon runs from the cecum to the rectosigmoid and is divided into ascending, transverse, descending, and sigmoid segments. Its wall forms semilunar folds and haustra (small pouches) created by the muscle layers, and the tube bends at the hepatic and splenic flexures. These structural features are normal, but they matter for visibility: folds create recesses and overhangs, bends change the camera angle, and variable distension changes the shape during the exam. As a result, some mucosal areas can be occluded from a forward-looking camera unless the endoscopist adjusts angulation, insufflation, and cleaning, or revisits a region during withdrawal. In practical terms, anatomy and “hidden corners” influence the success of complete inspection, especially for subtle or flat lesions tucked behind folds.

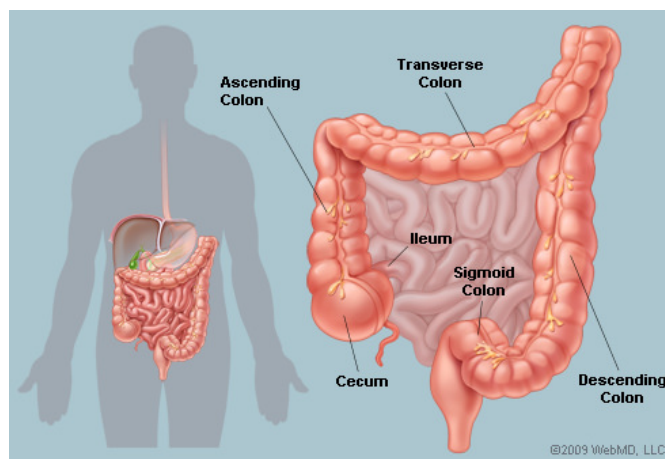


Figure 1.2: Colon anatomy³

³<https://michigansurgery.com/general-surgery/colon-surgery/>

1.1.4 Monocular endoscopy

A standard colonoscope is a monocular system: one camera sensor and a wide-angle objective at the tip, with the light source co-located next to the lens. The scope also has channels for air insufflation, water/irrigation, suction, and a working channel for tools like snares, forceps or injectors. The optical design favors near-field work: a large field of view and generous depth of field provide stable, continuous 2D video at high frame rate. The endoscopist advances to the cecum, then slowly withdraws while steering the tip and adjusting insufflation to unfold surfaces and improve visualization. From an imaging point of view, the setup has clear strengths: the scope is slim, the workflow is well established, and it delivers high-quality live video while allowing therapy through the same instrument. At the same time, the video is 2D only, so there is no depth information by default. The light sits next to the camera, which creates specular highlights on wet mucosa and changes the look of tissue as the scope moves. The view is also often occluded by folds, bends, fluids, or debris, and the colon itself changes shape with insufflation, peristalsis, and breathing. All together, these factors make some surfaces hard to inspect consistently and make it difficult to document with certainty what was actually seen versus what may have been missed during a routine exam.



Figure 1.3: Monocular endoscope setup⁴

⁴<https://medical.olympusamerica.com/products/evis-exera-iii-surgical-endoscopy>

1.2 Motivation of the thesis

A 3D model of the colon during or after colonoscopy can add information that the 2D video alone cannot provide. It gives depth, shape and continuity of the inspected surfaces, and it makes explicit which areas were seen and which were not.

A second motivation concerns the evaluation of examination quality. In clinical practice, the performance of a colonoscopy is usually assessed through standardized indicators such as the Adenoma Detection Rate (ADR), the Cecal Intubation Rate (CIR), and the average withdrawal time. Although these metrics are widely used and clinically meaningful, they remain indirect measures: they describe procedural success or overall outcomes but not the actual completeness of mucosal visualization. Even in optimal conditions, a relevant fraction of the surface can remain unobserved, and small or flat lesions may still go undetected. These limitations highlight that traditional quality metrics cannot fully capture the spatial completeness of an examination. For this reason, a reconstruction-based approach can also serve as an objective tool for performance assessment: by estimating the percentage of mucosal surface effectively visualized during the procedure, it becomes possible to quantify the thoroughness of the inspection in spatial terms and compare results across operators or sessions.

Third, follow-up after polyp treatment benefits from a stable spatial reference. With a 3D model from the index exam, a future colonoscopy can be aligned to the prior map, helping to relocate treated sites, track changes in scars or residual tissue, and compare subtle morphology over time.

Fourth, a 3D model can support pre-operative planning. Knowledge of the colon's curvature, flexures, length, and regions prone to looping can help anticipate difficult segments or incomplete exams. Imaging literature shows that anatomy and tortuosity are relevant to completion and strategy; mapping this geometry explicitly can inform planning for complex cases or repeat procedures.

Finally, reconstruction directly addresses the known limitations of 2D monocular imaging. White-light colonoscopy can miss lesions, especially small, flat or serrated ones, due to folds, bends and variable distension. Meta-analysis of tandem colonoscopies estimates an overall adenoma miss rate of about a quarter, underscoring the need for better ways to reason about what was and was not inspected. A 3D map is not a detector by itself, but it provides the spatial context needed to (i) quantify coverage, (ii) guide re-inspection during the same session or at follow-up, and (iii) normalize lesion locations for reporting and teaching.

1.3 Aim of the thesis

The main objective of this thesis is to develop and evaluate a computational framework capable of reconstructing a three-dimensional representation of the colon starting from standard monocular endoscopic videos. The purpose is to transform the conventional 2D visual information obtained during a colonoscopy into a coherent spatial model that captures both the geometry of the observed mucosal surface and the trajectory of the endoscope inside the lumen. The 3D reconstruction aims to provide a global spatial understanding of the colon from sequences of video frames, offering a complementary view to the real-time clinical image. By doing so, it becomes possible to identify which regions of the mucosa were visualized and which remained unobserved due to folds, occlusions, or limited camera coverage. This representation is designed not only for geometric reconstruction, but also for quantitative analysis. From the reconstructed model, it is possible to calculate metrics such as the percentage of unobserved surface, the distribution of missing regions, and the accuracy of the reconstructed geometry compared to reference data. These outputs can be used to objectively evaluate the completeness and quality of a colonoscopic examination. In summary, this thesis aims to bridge the gap between the limited spatial information of standard 2D colonoscopy, and the rich geometric understanding offered by 3D reconstruction, providing a foundation for more objective, spatially aware evaluation and documentation of colonoscopic procedures.

1.4 Contribution of the thesis

This thesis presents a series of methodological and experimental contributions aimed at improving spatial understanding and quantitative evaluation of colonoscopic procedures through 3D reconstruction and analysis. The main contributions are summarized below.

1.4.1 Development of a 3D reconstruction system

A complete pipeline was designed to reconstruct and visualize a three-dimensional model of the colon starting from monocular endoscopic videos. The system integrates RGB frames, depth maps, and camera poses to create a dense and anatomically coherent 3D surface. The reconstruction is based on a TSDF (Truncated Signed Distance Function) fusion approach, which merges frame-by-frame spatial information into a consistent volumetric representation.

1.4.2 Implementation of a mesh analysis framework for missing-region detection

A dedicated analysis module was developed to identify and quantify unobserved regions of the reconstructed surface. This system compares the reconstructed mesh with a closed version obtained using Poisson Surface Reconstruction. By computing geometric differences between the open and closed models, the method detects areas that were not covered by the camera’s field of view during the procedure. The analysis outputs several quantitative indicators, such as the total percentage of unobserved surface, the number and size of missing regions, and their spatial distribution along the colon. These metrics provide objective feedback on inspection completeness and procedural quality.

1.4.3 Comparative study of depth estimation models

A comprehensive comparison of state-of-the-art monocular depth estimation models from the literature was performed to identify the most effective network for colonoscopy scenes. Models such as Depth Anything, Video Depth Anything, and DepthPro were tested on synthetic and panthom endoscopic datasets. The comparison evaluated prediction accuracy, stability across frames, and compatibility with the subsequent 3D reconstruction step. This analysis highlighted the specific challenges of endoscopic imagery, such as specular reflections, low texture, and

rapid motion, and provided insights into model adaptation for this domain.

1.4.4 Acquisition of a real dataset combining video frames and camera poses

To validate the reconstruction approach on real data, a custom acquisition setup was created. An Olympus endoscope was used to record colonoscopy-like videos on a silicone colon phantom, while the camera pose was simultaneously tracked using the AURORA NDI electromagnetic tracking system for surgical instruments. This setup produced a dataset containing synchronized video frames and ground-truth camera trajectories, allowing realistic testing of reconstruction algorithms under controlled but physically plausible conditions.

Together, these contributions form an integrated framework that connects monocular 3D reconstruction, quantitative surface analysis, model benchmarking, and experimental data collection. The outcome is a reproducible and extensible system that supports both methodological research and potential clinical applications in colonoscopy quality assessment and spatial analysis.

Chapter 2

Background

2.1 3D reconstruction in medicine: overview and applications

Three-dimensional reconstruction is now a cross-disciplinary capability in medicine, supporting measurement, surgical planning, device design, augmented or mixed reality visualization, and 3D printing. Traditional volumetric methods like CT, MRI, and 3D ultrasound produce patient-specific models through segmentation and surface extraction that can be measured, simulated, or printed. In parallel, video-based approaches bring 3D into the operating room by estimating geometry directly from endoscopic or laparoscopic video. The following sections outline conventional tomographic techniques and then review video-based approaches, highlighting their principles and applications in clinical settings.

3D reconstruction already underpins many routine workflows. In colonoscopy, 3D mapping of the lumen helps document which mucosal regions were seen and which were missed; several systems now achieve real-time dense surface reconstruction that explicitly highlights unobserved areas. In orthopaedic and cranio-maxillofacial surgery, tomographic 3D models and prints assist with fracture reduction, implant fitting, and patient-specific surgical guides. Neurosurgical planning benefits from 3D reconstructions coupled with VR or MR visualization for complex cranial or skull-base procedures. Finally, stereo endoscopic systems have achieved dense surface reconstruction at clinical frame rates for navigation and local mapping, while video-based monocular pipelines are being explored to extend 3D awareness to flexible endoscopy. [1]

2.1.1 Traditional techniques

Computed Tomography (CT)

In CT-based reconstruction, the process starts with volumetric imaging acquired through a protocol tailored to the target anatomy, specifying parameters such as slice thickness, reconstruction kernel, and contrast phase. After acquisition, relevant tissues are segmented—either manually, semi-automatically, or with deep-learning tools—and the corresponding surfaces are extracted, typically using the Marching Cubes algorithm. The resulting meshes are then refined through smoothing and decimation before being exported for visualization, simulation, or 3D printing. As a well-established and quantitatively reliable technique, CT remains the reference standard for evaluating emerging three-dimensional reconstruction methods.

Magnetic Resonance Imaging (MRI)

MRI extends tomographic reconstruction to soft tissues with high contrast and no ionizing radiation. The process parallels CT—segmentation followed by surface extraction—but must handle motion artefacts, intensity inhomogeneity, and lower spatial resolution. Multi-contrast segmentation (T1/T2) is often used to delineate complex anatomy before meshing. The models support pre-operative planning and computational simulation, though longer acquisition times and voxel anisotropy can limit geometric accuracy.

Ultrasound (US)

3D ultrasound enables volumetric imaging without ionizing radiation, using either mechanically or robotically swept 2D probes, or matrix array transducers capable of acquiring volumetric data directly. During acquisition, each 2D frame is spatially registered—through probe tracking or known sweep geometry—and then integrated into a 3D voxel grid in a process known as scan conversion. Subsequent segmentation must account for typical ultrasound artefacts such as speckle noise and acoustic shadowing, yet reliable 3D contours can be reconstructed for clinical use. These models are routinely applied in obstetrics, cardiology, and image-guided interventions. Compared with CT or MRI, ultrasound-based reconstructions are more operator-dependent but provide real-time volumetric feedback directly at the bedside.

X-ray Tomography

Three-dimensional reconstruction from X-rays relies on acquiring multiple two-dimensional projections of the same object from different angles. Each image records the cumulative attenuation of the X-ray beam along its path, and by

combining these projections through mathematical inversion—typically filtered back-projection or iterative reconstruction—the internal 3D distribution of tissue density can be recovered. The resulting volumetric dataset can then be rendered or segmented to visualize anatomical structures. Although it involves ionizing radiation, X-ray-based reconstruction remains a cornerstone of medical imaging, forming the conceptual basis of computed tomography and other tomographic modalities.

2.1.2 Video-based techniques

Stereo endoscopy

Stereo endoscopes capture synchronized left–right views, producing metric depth through stereo correspondence without scale ambiguity. The pipeline includes calibration, rectification, disparity computation (classical or learned), and fusion of per-frame depth maps using TSDF or mesh integration. GPU-accelerated systems can achieve real-time dense reconstructions with millimetric accuracy, supporting navigation and surface mapping. While they provide robust local geometry, stereo devices require careful calibration and add hardware complexity.

Monocular endoscopy

Monocular endoscopes remain the standard in many clinical domains, such as gastroenterology, due to their compact design and integration of illumination at the tip. Reconstructing 3D geometry from their 2D video streams provides spatial context even in the absence of stereo imaging. Approaches to monocular 3D reconstruction can be broadly grouped into geometric, photometric, and learning-based families. Geometry-based methods, such as Structure-from-Motion and SLAM, detect and match features across frames to estimate camera poses through epipolar geometry; triangulation and bundle adjustment then recover scene structure, and dense surfaces are obtained through multi-view stereo or depth-from-motion before being fused into stable maps. Photometric methods exploit controlled lighting conditions at the endoscope tip, using shape-from-shading or photometric stereo to infer surface normals and depth from intensity variations, allowing sub-millimetric reconstruction of mucosal surfaces under calibrated illumination. More recently, learning-based and hybrid approaches employ deep networks to predict depth—and sometimes pose—directly from monocular frames, enforcing geometric and temporal consistency across sequences. These predictions can be fused into coherent 3D maps or combined with SLAM-based motion estimation to improve accuracy. Monocular systems still face intrinsic difficulties such as scale ambiguity, small baselines, repetitive texture, specular highlights, and soft-tissue deformation, which make correspondence and metric reconstruction unstable. Nonetheless, with proper

calibration, robust estimation, and drift correction, monocular reconstruction offers a practical path to spatial reasoning in endoscopy, leveraging standard clinical hardware without altering established procedures.

2.2 3D reconstruction techniques

2.2.1 Geometric multi-view methods

Geometric multi-view methods are image-based approaches that recover 3D structure and camera motion from overlapping images by exploiting the projective camera model and epipolar constraints to enforce cross-view consistency. They triangulate matched correspondences into 3D points and refine both structure and poses by minimizing reprojection error with bundle adjustment. The result is a calibrated set of camera poses and a sparse 3D scaffold that can be densified with multi-view stereo, providing a principled, geometry-first foundation for image-driven 3D reconstruction with well-studied accuracy–robustness trade-offs.

Structure from Motion (SfM)

Structure from Motion (SfM) is a fundamental geometric technique for reconstructing a camera’s trajectory and the 3D structure of a scene using multiple overlapping images. It works by detecting visual features, matching them across frames, estimating camera poses using epipolar relations, triangulating 3D points, and then applying bundle adjustment to refine both the camera parameters and the 3D points so that their reprojections align with the original images. [2] Two main workflows exist in practice: incremental systems that add images one at a time, and global systems that compute initial poses for all images via the view-graph before optimization. The open-source tool COLMAP [2] is one of the most widely used implementations of these ideas. In medical applications, SfM has been applied in settings where the anatomy remains relatively rigid over the captured frames, or where some anatomical features provide stable reference points. However, in endoscopic imaging—especially of the colon—SfM faces several challenges: monocular video lacks absolute scale, parallax is often limited, mucosal texture is repetitive or low, specular reflections are frequent, and organ deformation is common.[3] For colonoscopy, a recent feasibility study by 3D Reconstruction of the Human Colon from Capsule Endoscope Video [4] explores how sequences from wireless capsule endoscopy can be used to reconstruct whole sections of the human colon using SfM/SLAM-based pipelines. The authors employ a virtual graphics-based model of the gastrointestinal tract to provide ground-truth for geometric evaluation, demonstrating that while dense surface recovery is achievable under idealized conditions, accurate camera pose estimation remains challenging

due to scale drift, deformation, and image artifacts. In summary, SfM provides a robust, hardware-agnostic starting point for reconstructing geometry from images. Its strengths lie in its well-understood mathematical foundations and widespread availability of open-source tools. Yet in endoscopic settings—where scale is ambiguous, parallax is small, reflections litter the scene, and tissues deform—SfM alone often falls short. It must therefore be augmented with additional cues (such as loop closure, pose-graph optimization, learned priors or sensor fusion) to achieve 3D reconstructions that are clinically meaningful.

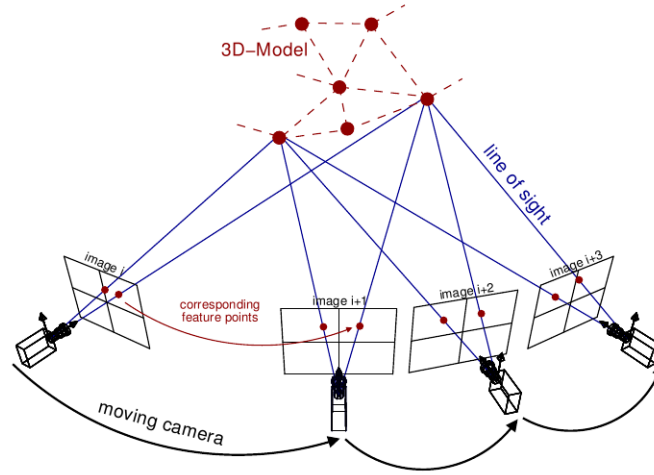


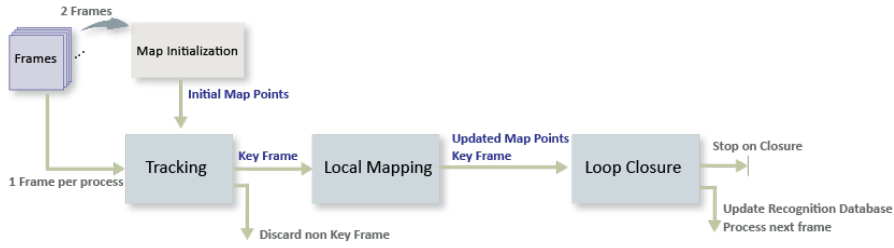
Figure 2.1: Structure from Motion principle¹

Simultaneous Localization and Mapping (SLAM)

Simultaneous Localization and Mapping (SLAM) is a process in which a moving camera or sensor builds a representation of an environment while simultaneously estimating its own pose within that evolving map. In the context of 3D reconstruction, SLAM does more than simply track movement: it incrementally accumulates spatial information from successive frames, fusing pose estimates and depth cues into a coherent 3D geometry of the scene. Over time, the system refines its map—typically a point cloud, mesh or volumetric model—while correcting drift and improving consistency through loop closures and global optimisation. [5] The input is a calibrated video (monocular in most flexible endoscopes; stereo/RGB-D

¹https://www.researchgate.net/figure/Structure-from-Motion-SfM-photogrammetric-principle-Source-Theia-sfmorg-2016_fig3_303824023

for some rigid scopes). A modern pipeline alternates tracking—estimating the current pose relative to a live map—and mapping, which integrates new frames; loop closure and pose-graph optimization correct drift. The output is a time-stamped 6-DoF trajectory plus a map ranging from sparse landmarks to dense surfaces via volumetric fusion or meshing. ORB-SLAM2 [6] exemplifies real-time tracking with relocalization and loop closure across monocular, stereo, and RGB-D inputs; dense surfaces are commonly produced downstream with TSDF-style fusion. Endoscopic imagery adds practical difficulties—specular highlights, repetitive/low texture, rolling shutter, and soft-tissue motion—so recent systems pair geometric tracking with learned representations and robust fusion. EndoGSLAM [7], for instance, uses a streamlined 3D Gaussian representation with differentiable rasterization to achieve real-time dense mapping and view synthesis, improving the trade-off between intraoperative availability and reconstruction quality. Illumination-aware conditioning further stabilizes tracking and mapping. Early dense visual SLAM for colonoscopy established that it is possible to reconstruct local mucosal surfaces online while explicitly leaving unobserved regions empty. RNNSLAM [8] formalized this idea by coupling a recurrent network that predicts scale-consistent depth and camera pose with a SLAM optimizer that fuses predictions into coherent surfaces and reduces drift; the system delivers real-time “chunk” reconstructions that make coverage gaps visible during the exam. Subsequent work broadened the design space. Multiple-map pipelines improved robustness over full procedures by favoring stronger features and GPU-accelerated matching: CudaSIFT-SLAM [9] reports sub-map merging and reliable relocalization across long, challenging sequences, mapping a substantially larger fraction of frames than feature-bag baselines. Topological formulations then addressed long-range place association despite deformation and low texture: ColonSLAM [10] links metric submaps into a global colon graph using deep place recognition and transformer-based matching, enabling whole-procedure maps on real data. Altogether, SLAM remains the most promising route to 3D reconstruction from routine endoscopic video—hence the focus of much recent work—but clinically robust deployment still contends with monocular scale ambiguity, limited parallax, repetitive/low mucosal texture, specular highlights, illumination changes, motion blur, and non-rigid tissue deformation; current surveys emphasize that progress hinges on hybrid designs that fuse geometric tracking with learned priors, illumination modeling, and deformation handling.[11]

Figure 2.2: SLAM principle²

2.2.2 Photometric methods

Photometric methods recover 3D shape by interpreting image intensities under a calibrated illumination/reflectance model, rather than relying on wide multi-view baselines. In endoscopy the light is near and often co-located with the camera, so brightness and multi-illumination cues become informative for monocular reconstruction when texture and parallax are limited. The inputs are standard endoscopic frames (or short sequences) plus photometric calibration of the camera and light(s) (e.g., positions/intensities and vignetting/specular handling). The outputs are dense local surface estimates—typically normals and depth—that can be fused across frames into meshes or volumetric maps, capturing high-frequency mucosal topography. Photometric methods recover 3D shape by interpreting image intensities under a calibrated illumination/reflectance model, rather than relying on wide multi-view baselines. In endoscopy the light is near and often co-located with the camera, so brightness and multi-illumination cues become informative for monocular reconstruction when texture and parallax are limited. The inputs are standard endoscopic frames (or short sequences) plus photometric calibration of the camera and light(s) (e.g., positions/intensities and vignetting/specular handling). The outputs are dense local surface estimates—typically normals and depth—that can be fused across frames into meshes or volumetric maps, capturing high-frequency mucosal topography. Foundational surveys detail how assumptions on reflectance and near-light geometry govern accuracy, motivating careful modeling for metric recovery. [12]

²<https://it.mathworks.com/help/vision/ug/monocular-visual-simultaneous-localization-and-mapping.html>

Shape from Shading (SfS)

Shape-from-Shading infers 3D surface shape from a single intensity image by explaining how brightness varies under a calibrated illumination and reflectance model. In practice, the inputs are a monocular image (or short sequence), camera intrinsics, and a photometric model of the light source(s)—often near and co-located with the camera in endoscopy—plus assumptions on surface reflectance (e.g., Lambertian with specular handling). The outputs are dense per-pixel surface normals and/or a depth map that can be integrated or fused across frames into meshes or volumetric reconstructions, providing high-frequency geometric detail where multi-view baselines are small. [13] Applications to colonoscopy. Recent work applies SfS to monocular colonoscopy frames to recover depth and local geometry of mucosa. Ruano et al.[14] train a learned SfS model on a large, realistic synthetic colonoscopy dataset and estimate depth maps from single frames, showing strong performance on synthetic benchmarks and qualitatively consistent depth on real procedures—depth that can then be used for 3D reconstruction tasks. Despite limitations such as specular reflections, shadows and deformation, SfS remains appealing for its hardware simplicity, real-time potential and dense per-frame output, which can later feed surface fusion or mapping pipelines.

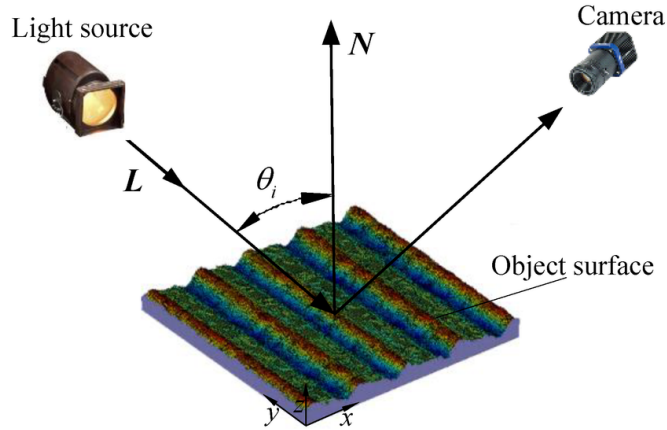


Figure 2.3: Shape from Shading principle³

Photometric Stereo (PS)

Photometric stereo recovers 3D surface shape by observing how image intensities change under varying illumination with (approximately) fixed viewpoint. In

³https://www.researchgate.net/figure/Shape-from-shading-SFS-model_fig1_331042184

practice, an endoscope sequentially activates tip-mounted LEDs (or uses rapid multiplexing), so each frame sees the same surface from the same camera but under a different light, with a calibrated illumination model—crucially, near-point lighting for endoscopy—one estimates per-pixel surface normals and albedo, then integrates the normals into depth and fuses results over frames into meshes or volumetric maps. [15] In practice, photometric methods for endoscopy take as input a calibrated stack of images from (approximately) a fixed viewpoint under varied, near-point illumination—i.e., camera intrinsics plus per-light calibration and basic reflectance assumptions—and produce dense per-pixel surface normals (and albedo) that are then integrated into depth and fused across frames into a coherent mesh or volumetric map. In monocular capsule endoscopy, Hao et al. [16] demonstrated that tip-mounted multi-LED photometric stereo can recover per-frame depth maps from single views, enabling 3D accumulation for navigation and mapping. In calibrated colonoscopy, Martínez Batlle et al. [17] exploited controlled near-light illumination to achieve the first in-vivo single-view photometric reconstruction of the human colon, introducing an in-place photometric calibration and reporting mean depth error under 3 mm on simulated studies with qualitative 3D on real sequences.

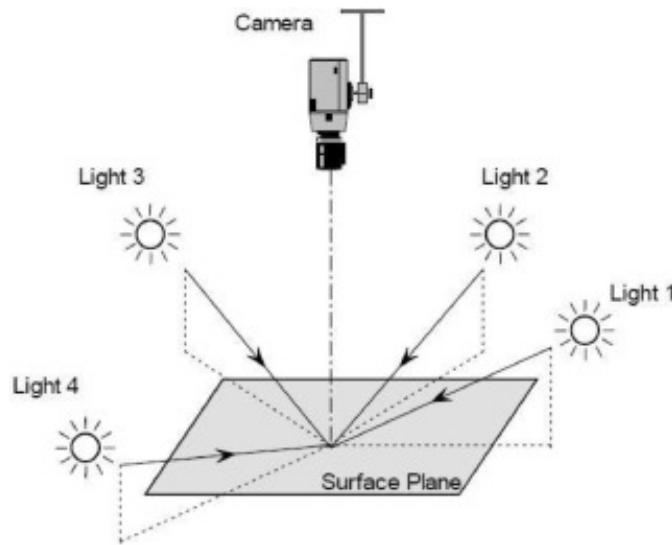


Figure 2.4: Photometric Stereo principle⁴

⁴<https://www.sciencedirect.com/science/article/pii/S129620742400270X>

2.2.3 Learning based methods

Learning-based models infer per-frame dense depth maps and 6-DoF camera poses directly from endoscopic video, replacing or augmenting classic geometric pipelines based on hand-crafted features and triangulation. Such networks are typically trained either with synthetic supervision (rendered depth/pose) or via self/weak supervision using photometric, geometric and temporal consistency across frames. At test time the network returns (i) a dense depth map aligned to each input frame and (ii) a timestamped pose trajectory; these outputs can be fused (e.g., via TSDF/mesh) into a navigable surface or used for coverage analytics. Surveys in endoscopic 3D confirm that learning-based methods are especially promising under low-texture, specular highlights, and small-baseline conditions typical of GI video.[18]

Depth estimation models

A depth map assigns to each image pixel an estimate of the distance to the visible surface along the camera ray; in practice it can be relative (up to scale) or metric (with absolute scale), and is later fused across frames into meshes or TSDF volumes for 3D reconstruction.

Neural models take single frames or short clips and regress dense depth using convolutional/Transformer backbones. Supervision comes either from synthetic data with ground-truth depth/pose, or from self/weak supervision that matches neighbouring frames photometrically and enforces geometric or temporal consistency—useful when real ground truth is unavailable in endoscopy.

Beyond task-specific networks, recent “foundation” models learn from massive, generic image corpora and transfer surprisingly well to medical video.

Depth Anything [19] and Depth Anything V2 [20] are trained at scale (tens of millions of unlabeled images with pseudo-labels), yielding strong zero-shot single-image depth and providing a practical prior to fine-tune for clinical domains; Video Depth Anything [21] extends this idea to long videos with a lightweight spatio-temporal head for temporally consistent depth. Depth Pro [22] targets sharp metric depth in a zero-shot setting, producing high-resolution maps without camera metadata. Recent depth-estimation pipelines tailor learning to GI constraints—small baselines, specularities, and deformation—while aiming for video-consistent geometry. Xu et al.[23] introduce a self-supervised framework with generative latent priors: a learned latent bank regularizes the network so depth (and pose) stay stable under illumination change and low texture; evaluated on synthetic and mixed endoscopic datasets, it improves depth quality without ground-truth labels, making it attractive for colon video where supervision is scarce.

To address long-sequence temporal consistency, ColonCrafter [24] formulates monocular colon depth as a diffusion-based conditional generation task, training

on synthetic colonoscopy sequences and applying a lightweight style-transfer step when deployed on real videos; the result is per-frame depth that remains consistent across time and is well-suited to TSDF/mesh fusion for 3D reconstruction.

Finally, geometry-aware designs from monocular endoscopy transfer well to the colon: Yang et al. [25] impose normal consistency and robust photometric terms to stabilize predictions under non-Lambertian reflectance and exposure changes—principles that directly strengthen colon depth estimation when paired with synthetic pretraining or self-supervision. In conclusion, learning-based depth estimation delivers a viable approach for extracting dense 3D geometry from endoscopic video, though key challenges remain—most notably achieving metric scale, ensuring device and patient generalization, handling lighting variations, and compensating for tissue deformation—which must be addressed for confident clinical application.

Pose estimation models

In computer vision and endoscopy, a camera’s pose refers to its position and orientation in three-dimensional space.

It is typically represented as a rigid transformation $T = [R \mid t]$, where R is a 3×3 rotation matrix (orientation) and t is a 3-vector (translation).

Accurate pose estimation is essential for aligning frames, reconstructing scenes, and tracking the endoscope path within the anatomy.

Video-based pose estimation pipelines generally operate in one of two fashions. In geometric approaches, features like keypoints or descriptors are detected and matched across frames or views; the camera motion is estimated via the essential or fundamental matrix, triangulation gives sparse 3D points, and global optimisation refines the trajectory. In learning-based systems, neural networks regress the 6-DoF transformation between frames (or with respect to a global map) or jointly predict pose and depth in a structure-from-motion learning paradigm.

These methods may rely on synthetic supervision, self/weak supervision, or a hybrid of geometry + learning.

In all cases, the input is endoscopic video, and the output is a timestamped pose trajectory.

In colonoscopy, estimating the pose of the endoscope is crucial to determine which parts of the colon mucosa have been inspected, to locate lesions or polyps in a spatial reference, and to enable instrument navigation and tracking over time. For example, Rau et al. [26] introduced the SimCol synthetic dataset for colonoscopy and proposed a bimodal pose-regression network trained on simulated colonoscope trajectories with ground-truth poses; the approach generalised to real colonoscopy sequences and outperformed preceding unimodal networks. Xu et al. [23] propose a self-supervised monocular depth + pose estimation framework

that incorporates a Generative Latent Bank and a Variational Autoencoder to regularize both depth and camera motion in GI videos, demonstrating improved pose accuracy on colonoscopy data. On the VO side, ColVO [27] targets procedure-length robustness in colonoscopy by coupling depth and pose and enforcing light-consistent calibration to handle the moving, near-coaxial light source. The result is more stable trajectories under brightness fluctuations and small baselines typical of the colon, with demonstrations on real colonoscopy videos.

Overall, recent research shows that reliable pose estimation in colonoscopy is achieved by integrating geometric tracking with learned depth and motion priors, improving stability across long, low-texture sequences and supporting accurate spatial reconstruction of the examined mucosa.

2.3 3D Representation and Visualization

A 3D representation is a digital model of anatomy designed for visualization, measurement and interaction. In endoscopic and surgical workflows, this stage converts camera poses and depth or CT/MR volumes into models that facilitate virtual inspection, quantitative assessment and navigation. Recent reviews of endoscopic 3D reconstruction emphasise that the choice of representation critically influences accuracy, rendering speed, memory footprint and ultimately what clinical tasks the system can support. In clinical practice, anatomy is typically visualised either as surfaces or as volumes. These visualisation modes are standard in medical workstations and open platforms such as 3D Slicer, enabling routine planning and review. In the following subsections we introduce the main representation families considered in this work.

2.3.1 Point clouds

A point cloud is a finite set of 3D samples

$$\mathcal{P} = \{(\mathbf{x}_i, c_i, \mathbf{n}_i)\}_i,$$

where $\mathbf{x}_i \in \mathbb{R}^3$ denotes the 3D position of the i -th sample, c_i its colour and \mathbf{n}_i an estimated surface normal. It is an explicit, sample-based representation without connectivity—fast to compute and visualise—and usually the first stable output after depth and pose estimation.

Given camera intrinsics \mathbf{K} , a per-pixel depth map $D_t(\mathbf{u})$ at frame t and pose $\mathbf{T}_t \in \text{SE}(3)$, a pixel $\mathbf{u} = (u_x, u_y)$ is back-projected into camera coordinates as

$$\mathbf{x}_{\text{cam}} = D_t(\mathbf{u}) \mathbf{K}^{-1} \begin{bmatrix} u_x \\ u_y \\ 1 \end{bmatrix}, \quad (2.1)$$

and transformed into the world frame by

$$\mathbf{x} = \mathbf{T}_t \mathbf{x}_{\text{cam}}. \quad (2.2)$$

Accumulating these samples across frames produces a dense, coloured point cloud. Local normals \mathbf{n}_i are commonly estimated via PCA on a k -nearest-neighbour neighbourhood; voxel-grid downsampling and statistical outlier removal are typical preprocessing steps before meshing or fusion.

The raw cloud typically exhibits uneven density and noise; voxel-grid downsampling regularizes point spacing, while statistical outlier removal eliminates isolated samples. When a watertight surface is needed, the oriented cloud can be converted

into a mesh using Poisson surface reconstruction, which fits an implicit function globally and extracts a triangle surface robust to moderate noise and missing regions. These steps transform scattered depth samples into a coherent 3D model suitable for visualization, measurement, or volumetric fusion in clinical endoscopy.

2.3.2 Polygonal meshes

A polygonal mesh represents a 3D surface as a collection of vertices, edges, and faces:

$$\mathcal{M} = (V, E, F),$$

where $V = \{\mathbf{v}_i \in \mathbb{R}^3\}$ are vertex coordinates, E the edge set, and F the set of polygonal (typically triangular) faces. Unlike point clouds, meshes encode topological connectivity, allowing continuous surface representation, curvature estimation, and physically consistent rendering.

Meshes are often reconstructed from oriented point clouds through algorithms that infer surface continuity. The most widely used is Poisson surface reconstruction, which estimates an implicit scalar field $f(\mathbf{x})$ whose gradient approximates the vector field of point normals and then extracts an isosurface:

$$\nabla^2 f = \nabla \cdot \mathbf{n}, \quad S = \{\mathbf{x} \in \mathbb{R}^3 \mid f(\mathbf{x}) = \text{iso}\}. \quad (2.3)$$

The resulting mesh is watertight and globally smooth, making it robust to moderate noise and outliers. For visualization and quantitative analysis, the mesh can be simplified through decimation, smoothed via Laplacian filtering, and textured by projecting color information from input images onto surface vertices. In endoscopic 3D reconstruction, polygonal meshes provide a compact, anatomically meaningful model of the mucosa, supporting measurements, navigation, and photorealistic rendering for clinical review.

2.3.3 Truncated Signed Distance Fields (TSDF)

A Truncated Signed Distance Function (TSDF) is an implicit volumetric representation of a surface. Instead of explicitly storing polygons, the scene is encoded as a scalar field $\phi(\mathbf{x})$ defined over a 3D voxel grid, where each voxel stores the signed distance to the nearest surface point. Distances are truncated to a narrow band $[-\tau, +\tau]$ around the surface to improve robustness and computational efficiency. The anatomical surface is represented by the zero level set of this field,

$$S = \{\mathbf{x} \in \mathbb{R}^3 \mid \phi(\mathbf{x}) = 0\}, \quad (2.4)$$

from which a mesh can be extracted for visualization or measurement.

Given the camera intrinsics \mathbf{K} , per-frame pose $\mathbf{T}_t \in \text{SE}(3)$, depth map $D_t(\mathbf{u})$, and TSDF parameters (voxel size, truncation distance τ , and weights), each pixel \mathbf{u} contributes a signed distance observation d_{new} along its viewing ray. The field is updated incrementally using a weighted average:

$$\phi_{\text{new}} = \frac{w_{\text{old}} \phi_{\text{old}} + w_{\text{obs}} d_{\text{new}}}{w_{\text{old}} + w_{\text{obs}}}, \quad w_{\text{new}} = \min(w_{\text{old}} + w_{\text{obs}}, w_{\text{max}}). \quad (2.5)$$

Once fused, surface normals can be estimated from the TSDF gradient,

$$\mathbf{n}(\mathbf{x}) = \nabla \phi(\mathbf{x}), \quad (2.6)$$

and the surface mesh is extracted using Marching Cubes at the zero-crossing $\phi(\mathbf{x}) = 0$. This volumetric fusion principle—first introduced by Curless and Levoy—remains the basis of many dense reconstruction systems. In colonoscopy, TSDF fusion combines per-frame depth and 6-DoF poses within a narrow band around the visible mucosa to yield smooth, watertight reconstructions. For example, Liu et al., [28] propose a sparse-to-dense pipeline for colonoscopic scenes where refined depth is fused via a TSDF stage to produce a dense 3D colon model. In the broader endoscopic literature, pipelines that estimate depth and pose from monocular video often perform volumetric fusion with TSDF to obtain a watertight mesh, reinforcing TSDF’s suitability for dense, real-time medical reconstruction [29].

2.3.4 Neural Radiance Fields (NeRF)

A Neural Radiance Field (NeRF) is a continuous, implicit scene representation parameterized by a neural network. Instead of discrete geometry, a function F_θ maps a 3D location and a viewing direction to a volume density and view-dependent color; novel views are rendered via differentiable volumetric ray integration, and the network is trained to reproduce a set of posed RGB images [30].

Inputs are calibrated intrinsics, camera poses, and corresponding RGB frames (optionally depth or scale priors). The output is a trained radiance field that enables photorealistic novel-view rendering; depth and normals can be derived from the rendering integral or from SDF-based variants.

NeRF defines

$$F_\theta : (\mathbf{x}, \mathbf{d}) \mapsto (\sigma, \mathbf{c}), \quad (2.7)$$

with $\mathbf{x} \in \mathbb{R}^3$ a 3D point, $\mathbf{d} \in \mathbb{S}^2$ a viewing direction, $\sigma \geq 0$ the volume density, and $\mathbf{c} \in \mathbb{R}^3$ the emitted color. For a camera ray $r(t) = \mathbf{o} + t\mathbf{d}_{\text{cam}}$, discrete samples $\{\mathbf{x}_i\}$ along the ray yield an approximate pixel color

$$C(r) = \sum_i T_i (1 - e^{-\sigma_i \Delta_i}) \mathbf{c}_i, \quad T_i = \exp \left(- \sum_{j < i} \sigma_j \Delta_j \right). \quad (2.8)$$

Training minimizes a photometric loss

$$L(\theta) = \sum_k \sum_{r \in I_k} \|C_\theta(r) - I_k(r)\|_2^2 + R(\theta), \quad (2.9)$$

with gradients back-propagated through the differentiable renderer. A depth estimate can be obtained as the expected termination distance

$$D(r) = \sum_i T_i (1 - e^{-\sigma_i \Delta_i}) t_i. \quad (2.10)$$

Weak depth/scale priors are often introduced to mitigate scale ambiguity and sparse viewpoints, conditions common in endoscopic data.

Endoscopic adaptations address specularities, sparse coverage, and tissue non-rigidity. For colonoscopy, ColonNeRF [31] reconstructs long procedures by dividing the colon into local segments and performing pose densification to stabilize training, improving texture realism and geometric consistency on synthetic, phantom, and clinical data. In surgical endoscopy, Efficient EndoNeRF [32] streamlines the NeRF pipeline for deformable scenes, targeting faster convergence and high-fidelity reconstructions suitable for simulation and training.

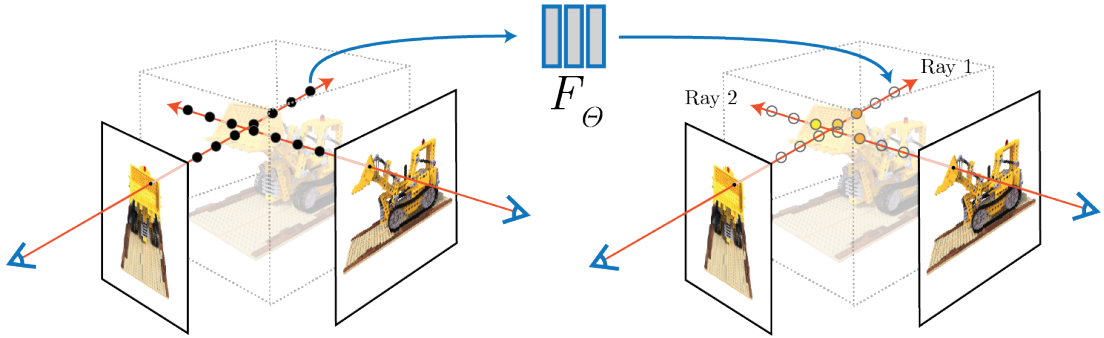


Figure 2.5: NeRF principle⁵

⁵<https://www.matthewtancik.com/nerf>

2.3.5 3D Gaussian Splatting (3DGS)

3DGS represents a scene with a set of anisotropic Gaussian primitives

$$\mathcal{G} = \{G_i\}, \quad G_i = (\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \mathbf{c}_i, \alpha_i),$$

where $\boldsymbol{\mu}_i \in \mathbb{R}^3$ (position), $\boldsymbol{\Sigma}_i \in \mathbb{R}^{3 \times 3}$ (covariance), $\mathbf{c}_i \in \mathbb{R}^3$ (color) and $\alpha_i \in [0, 1]$ (opacity). Each Gaussian projects to the image plane as a 2D ellipse and contributes by visibility-aware splatting with alpha compositing:

$$C(p) = \sum_k T_k \alpha_k \mathbf{c}_k, \quad T_k = \prod_{j < k} (1 - \alpha_j). \quad (2.11)$$

This screen-space formulation (instead of volumetric ray marching) enables real-time training and rendering for novel-view synthesis.

Inputs are camera intrinsics, posed RGB images, and typically an SfM/SLAM point seed for initialization; optional depth/segmentation priors guide geometry and dynamics. The output is an optimized set of Gaussians \mathcal{G} supporting photorealistic real-time rendering; exports include dense colored point clouds or meshes (via post-processing) when needed.

Parameters $\{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \alpha_i, \mathbf{c}_i\}$ are refined by minimizing a photometric objective over all frames, e.g.,

$$L(\Theta) = \sum_k \sum_{p \in I_k} \|C_{\Theta}(p) - I_k(p)\|^2 + R(\Theta), \quad (2.12)$$

with regularizers R for covariance conditioning and density control; gradients are back-propagated through the differentiable splat renderer.

Single-view deformable scenes: EndoGS [33] adapts 3DGS to deformable endoscopic tissue by introducing a deformation field, depth-guided supervision and spatio-temporal masks to handle tool occlusions from a single viewpoint, yielding high-quality reconstructions on robotic-surgery videos. EndoGaussian [34] focuses on speed and robustness in dynamic scenes via holistic Gaussian initialization and spatio-temporal Gaussian tracking, reporting ~ 195 FPS rendering and < 2 min per-scene training on public datasets—indicative of intraoperative potential. Endo-2DTAM [35] integrates 2D/3D Gaussian splatting into a real-time SLAM system with surface-normal-aware tracking/mapping and a BA module, improving geometric accuracy and dense reconstruction quality in endoscopic videos.

Together, these works position 3DGS as both a high-fidelity renderer and a geometric backbone for real-time, photorealistic endoscopic reconstruction, even under sparse views and tissue motion.

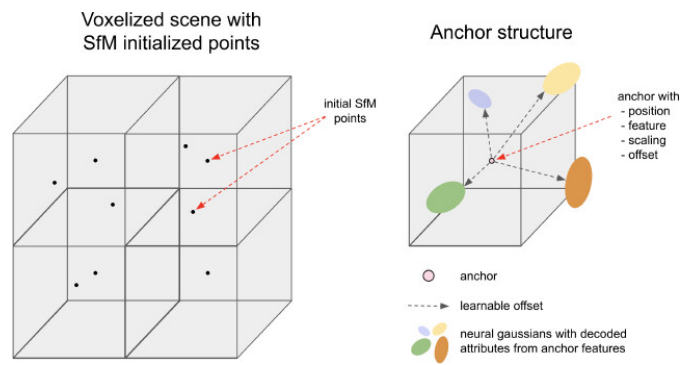


Figure 2.6: Gaussian Splatting principle⁶

⁶<https://onlinelibrary.wiley.com/doi/10.1111/cgf.70078>

2.4 Missing regions identification

2.4.1 Coverage in CT colonography

The notion of "missing regions" originates in CT colonography (CTC), where a complete 3D colon surface is available and the task is to quantify how much of that surface becomes visible along a virtual fly-through. Early CTC systems computed surface-coverage maps and flagged missed patches above a size threshold, establishing both the vocabulary (coverage, unseen areas) and the visualization patterns (color overlays, unfolded views) later adopted in optical workflows. They also explored view strategies (retrograde and antegrade review) to reduce blind spots [36].

2.4.2 Reconstruction-based coverage

With real-time reconstruction from colonoscopy video, missing regions become explicit in 3D: surveyed mucosa appears in the model, while unobserved areas remain absent. A landmark RNN-SLAM pipeline [37] predicts scale-consistent depth and camera pose from monocular frames and fuses them online, so coverage gaps emerge as literal holes in the reconstructed surface; the extended report demonstrates interactive performance on clinical sequences.

2.4.3 Real-time guidance toward unseen mucosa

Building on coverage estimation, navigation aids now provide actionable feedback. ColNav [38] couples a real-time unfolded representation with a local indicator that directs the endoscopist toward un-inspected regions; experiments report higher polyp recall and strong agreement with physicians' coverage assessments, translating geometric awareness into practical guidance at the scope.

2.4.4 Quantifying coverage on reconstructed surfaces

With a reconstructed surface or point cloud in hand, coverage becomes a measurable quantity rather than a visual impression. A representative method operates directly on the colon point cloud: it partitions the geometry into anatomically meaningful segments and then estimates per-segment coverage by inferring the area of unobserved regions. On synthetic and CT-derived segments it reports mean absolute errors of approximately 3–6%, with qualitative agreement on reconstructions from real procedures [39].

2.5 Related works for colon 3D reconstruction

This chapter reviews a set of representative, high-impact methods for monocular colon 3D reconstruction, chosen for their reliability, citation impact, and practical relevance.

2.5.1 RNN-SLAM

RNNSLAM [8, 40] proposes a deep-learning-driven dense SLAM system to reconstruct the colon in real time and explicitly visualize missing regions. A recurrent network predicts scale-consistent depth maps and 6-DoF poses from monocular colonoscopy frames; these predictions initialize a photometric SLAM back-end with local window optimization to reduce drift. Keyframe depths are then fused into a textured surface mesh, so unsurveyed areas remain unreconstructed and appear as gaps—providing actionable coverage feedback to the endoscopist. Evaluations report lower absolute pose error than baselines (DSO and standalone RNN-DP) across 12 colonoscopic sequences, with qualitative reconstructions shown at interactive rates. Data include real clinical colonoscopy videos (monocular RGB) and virtual colonoscopic images used for depth comparisons.

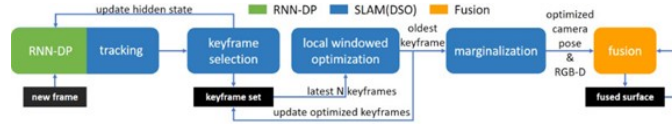


Figure 2.7: Pipeline of RNN-SLAM

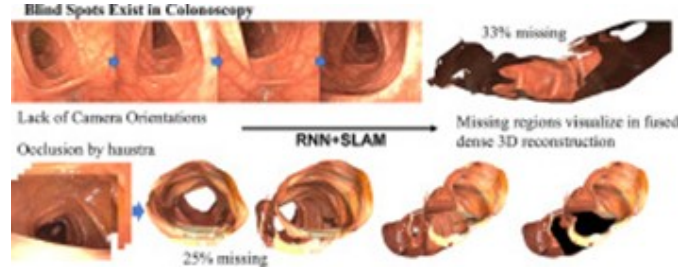


Figure 2.8: 3D reconstruction results of RNN-SLAM

2.5.2 ColVO

ColVO [27] is a deep learning-based visual odometry framework tailored to colonoscopy that jointly estimates depth and 6-DoF pose from monocular video.

Its design couples the two tasks via a Deep-Couple Depth-Pose (DCDP) strategy—enforcing geometric projection consistency between consecutive frames—and stabilizes tracking under moving, near-coaxial illumination with a Light-Consistent Calibration (LCC) module. The system alternates frame-to-frame motion estimation with depth refinement and produces a time-stamped trajectory together with per-frame depth that can drive real-time 3D mapping and lesion localization. Reported experiments show accuracy gains over state-of-the-art VO/SLAM baselines and demonstrate two applications: immediate 3D polyp localization and whole-procedure 3D reconstruction from monocular RGB.

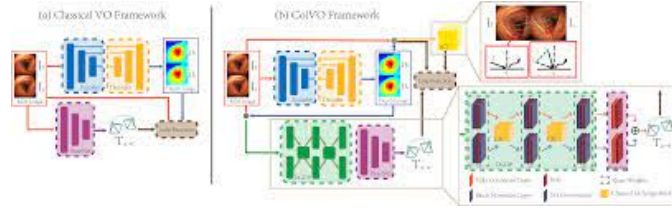


Figure 2.9: Pipeline of ColVO

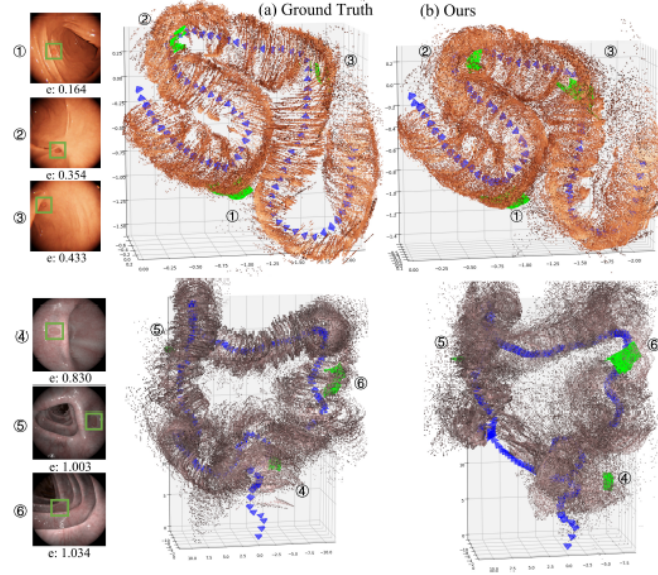


Figure 2.10: 3D reconstruction results of ColVO

2.5.3 Endo2DTAM

Endo-2DTAM [35] is a real-time endoscopic SLAM system that replaces volumetric ray marching with a 2D Gaussian splatting representation and couples it with a surface-normal-aware tracking-mapping-BA pipeline. Tracking blends point-to-point and point-to-plane distances to stabilise pose, while mapping enforces normal consistency and a depth-distortion term to improve geometric fidelity; a pose-consistent keyframe policy maintains coherent sampling over long sequences. The system delivers geometrically accurate, visually consistent reconstructions with real-time rendering and reports low depth error on public endoscopic datasets.

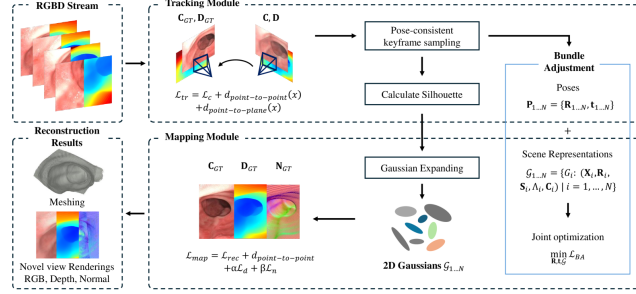


Figure 2.11: Pipeline of Endo2DTAM

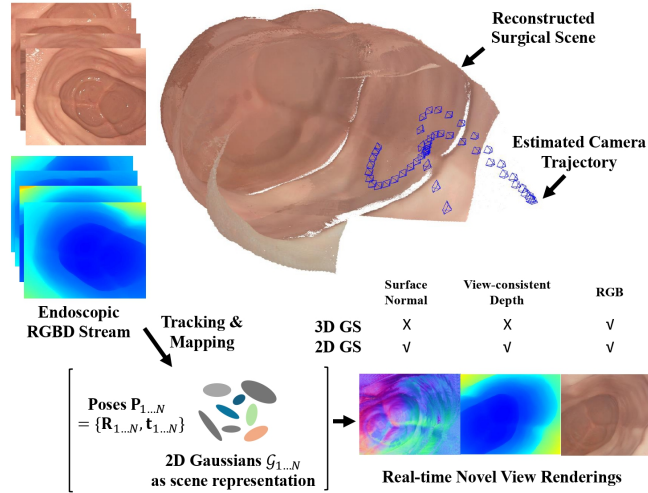


Figure 2.12: 3D reconstruction results of Endo2DTAM

2.5.4 EndoGSLAM

EndoGSLAM [7] is a dense SLAM framework for endoscopic surgeries that combines streamlined 3D Gaussian representation with differentiable rasterization for efficient camera tracking and tissue reconstruction. It achieves real-time dense mapping at over 100 fps while refining camera poses and expanding its 3D model to cover unseen areas. Evaluations on clinical videos demonstrate EndoGSLAM’s superior balance of reconstruction quality and intraoperative speed compared to existing SLAM methods, supporting real-time surgical navigation and visualization.

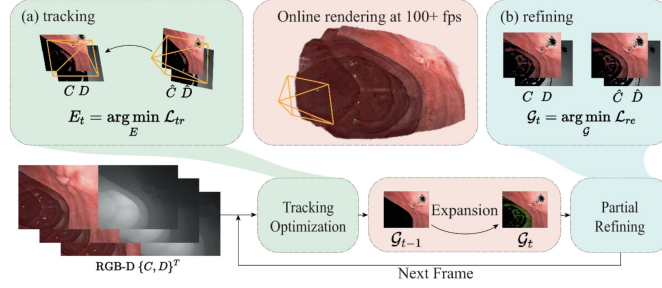


Figure 2.13: Pipeline of EndoGSLAM

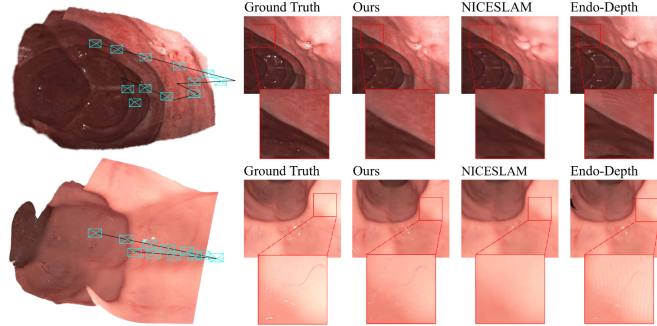


Figure 2.14: 3D reconstruction results of EndoGSLAM

2.5.5 C³Fusion

C³Fusion [41] is a deep-learning-based dense SLAM framework designed specifically for colonoscopy, aiming to achieve consistent depth and pose estimation through contrastive feature fusion. The system couples a contrastive encoder-decoder architecture with a differentiable fusion module that aligns depth and pose predictions from consecutive frames, enforcing geometric consistency across time. Unlike traditional photometric SLAM, C³Fusion learns feature correspondences directly in latent space, improving robustness to specular highlights, low texture, and deformation

typical of endoscopic scenes. The fused predictions are integrated into a volumetric 3D map that preserves global structure while reducing drift. Experiments on synthetic and clinical colonoscopy datasets demonstrate improved reconstruction accuracy and temporal stability over prior self-supervised and geometric baselines.

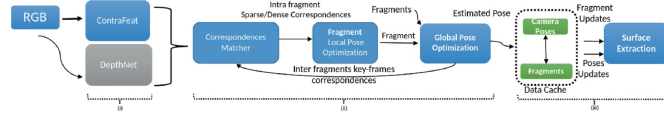


Figure 2.15: Pipeline of C³Fusion

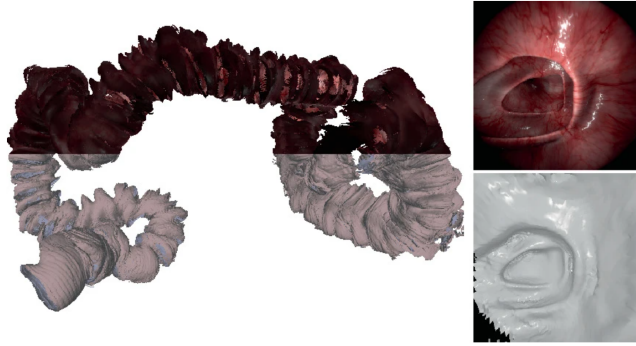


Figure 2.16: 3D reconstruction results of C³Fusion

2.5.6 Summary

Table 2.1 provides a comparative overview of the methods presented in the preceding sections, summarizing their key characteristics in terms of dataset, reconstruction approach, visualization strategy, and operational capabilities. The abbreviation **RT** indicates whether the system operates in **real-time**, while **MR** denotes the ability to explicitly identify and visualize **missing regions** in the reconstructed surface.

Table 2.1: Comparison of 3D reconstruction methods of the colonic surface.

Model	Dataset	3D Reconstruction Method	3D Visualization	RT	MR
RNNSLAM	Clinical colonoscopy videos (UNC)	Deep learning-driven dense SLAM	Textured mesh; Unfolded coverage map	Y	Y
CoIVO	Colonoscopy Simulation Dataset, VR-Caps	Coupled depth-pose VO (DCDP)	Dense Point Cloud (unspecified type)	N	N
Endo-2DTAM	C3VD	RGB-D SLAM with 2D Gaussian Splatting	Gaussian splatting novel view rendering	Y	N
EndoGSLAM	C3VD	SLAM with Gaussian Splatting + diff. rendering	Real time Gaussian splatting rendering	Y	N
C³ Fusion	Synthetic simulator ; Colon10K	Volumetric TSDF fusion + Marching Cubes	Textured mesh with re-rendered views	N	N

Chapter 3

Materials and Methods

3.1 Datasets

3.1.1 SimCol3D

SimCol3D¹ is a purpose-built dataset for studying depth, camera pose, and 3D reconstruction in colonoscopy. The synthetic component starts from three CT-derived colon anatomies that are converted into high-fidelity meshes and rendered with endoscopic optics and lighting (wide-angle lens, near-field illumination, wet mucosa, realistic camera motion). Each rendered frame is paired with the information needed for spatial reasoning: RGB imagery, per-pixel depth, full 6-DoF camera pose, and calibrated intrinsics. The dataset is organized by anatomy and trajectory: Synthetic Colon I and Synthetic Colon II each include 15 virtual colonoscopy trajectories (with dedicated training and test splits), while Synthetic Colon III provides 3 test trajectories intended for evaluation only. This structure makes it possible to train on diverse, labeled sequences and then test generalization on a distinct anatomy. In this thesis, SimCol3D is used to train and tests depth models and a supervised pose estimator, to supply RGB + depth + pose to the TSDF fusion pipeline, and to run ablation studies and missing-region analyses under consistent geometry.

¹<https://github.com/anitarau/simcol>



Figure 3.1: 3D models of the three synthetic colons the SimCol3D dataset. (SyntheticColon_I, SyntheticColon_II, SyntheticColon_III)

3.1.2 C3VD

C3VD (Colonoscopy 3D Video Dataset) ² consists of short real-looking colonoscopy sequences recorded with a clinical HD colonoscope while imaging high-fidelity silicone colon segments. Each video frame is registered to a known 3D colon model using a multimodal 2D–3D registration pipeline; this yields per-frame geometric labels derived from the aligned model. As released, the dataset includes 22 short, registered videos (10k labeled frames) and a small set of screening-style videos with paired ground-truth poses. For each labeled frame, C3VD provides RGB, per-pixel depth, surface normals, optical flow, occlusion masks, 6-DoF camera pose, coverage maps, and the corresponding 3D surface models. Since the videos cover short tracts-and not complete colonoscopies-C3VD is used to validate the reconstruction system by comparing TSDF-based meshes (built from RGB with predicted depth and pose) against the ground-truth 3D models provided in the dataset. For each short-tract sequence, the reconstructed mesh is aligned to the corresponding GT surface and evaluated with symmetric Chamfer-L1 distance and mesh-to-mesh surface overlap.

²<https://durrlab.github.io/C3VD/>

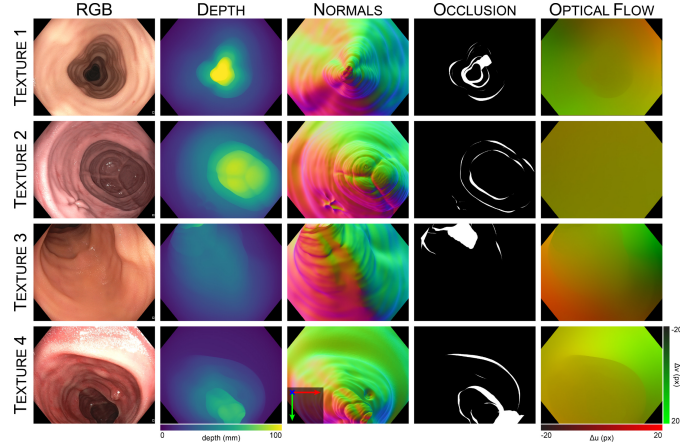


Figure 3.2: C3VD dataset

3.2 TSDF-module

This system builds a dense 3D model of the colon from per-frame depth maps, camera poses, and camera intrinsics. It fuses the sequence into a volumetric TSDF (truncated signed-distance field) and then extracts either a triangle mesh or a colored point cloud for visualization.

3.2.1 Inputs

RGB frames

- Dimensions: 475×475 pixels.
- Channels: 3-channel RGB PNG, 8-bit per channel (uint8).

Depth Maps

- Dimensions: 475×475 pixels.
- Channels : 1-channel grayscale PNG, 16-bit unsigned (uint16)

Camera Poses

For each frame i there are two synchronized lines from two different files :

SavedPosition.txt - Translation They are the camera position coordinates in the chosen world/reference frame (units must be consistent with the reconstruction, e.g., mm).

$$\mathbf{t}_i = \begin{bmatrix} t_x & t_y & t_z \end{bmatrix}^\top. \quad (3.1)$$

SavedRotationQuaternion.txt - Orientation Four floats (unit quaternion) in the order $q_x \ q_y \ q_z \ q_w$. This is the camera orientation quaternion

$$\mathbf{q}_i = (q_x, q_y, q_z, q_w)$$

for the same frame i . The quaternion should be normalized: $\|\mathbf{q}_i\| = 1$.

From Quaternion the Rotation matrix is obtained as follows: Given $\mathbf{q} = (q_x, q_y, q_z, q_w)$ with $\|\mathbf{q}\| = 1$, the 3×3 rotation matrix $R(\mathbf{q})$ is

$$R(\mathbf{q}) = \begin{pmatrix} 1 - 2(q_y^2 + q_z^2) & 2(q_x q_y - q_z q_w) & 2(q_x q_z + q_y q_w) \\ 2(q_x q_y + q_z q_w) & 1 - 2(q_x^2 + q_z^2) & 2(q_y q_z - q_x q_w) \\ 2(q_x q_z - q_y q_w) & 2(q_y q_z + q_x q_w) & 1 - 2(q_x^2 + q_y^2) \end{pmatrix}. \quad (3.2)$$

Properties: $R^\top R = I$ (orthonormal) and $\det R = +1$.

The **camera to world pose matrix** is then obtained as follows:

$$T_i^{(\text{cam} \rightarrow \text{world})} = \begin{bmatrix} R(\mathbf{q}_i) & \mathbf{t}_i \\ \mathbf{0}^\top & 1 \end{bmatrix}. \quad (3.3)$$

Camera Intrinsics

Format: Each frame has 9 whitespace-separated floats forming the row-major 3×3 pinhole matrix K :

$$K = \begin{pmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{pmatrix} \quad (3.4)$$

which maps normalized camera coordinates to pixel coordinates.

- f_x, f_y — Effective focal lengths (in pixels)
 - They scale horizontal and vertical directions:

$$f_x = \frac{f}{p_x}, \quad f_y = \frac{f}{p_y}, \quad (3.5)$$

where f is the physical focal length and p_x, p_y are pixel pitches. Larger f_x, f_y result in a narrower field of view.

- c_x, c_y — Principal point (in pixels)
 - The pixel coordinates where the optical axis intersects the image plane. Often near the image center but not exactly.
- s — Skew (pixel axis non-orthogonality)
 - Zero for most modern sensors ($s = 0$) with square pixels and orthogonal axes. If nonzero, horizontal and vertical pixel axes are slightly sheared.
- **Bottom row** $\begin{bmatrix} 0 & 0 & 1 \end{bmatrix}$ — Homogeneous scaling
 - Ensures the mapping works in homogeneous coordinates.

3.2.2 Data processing

Before the volumetric fusion begins, two key preprocessing steps are applied to reduce computational cost and standardize units:

Resizing

The depth (and optional RGB) images are downsampled (e.g., from their original resolution to 237 x 237 pixels) in order to reduce the voxel count and memory footprint in the TSDF. This resizing strategy reduces computational load while preserving adequate surface detail for the colon model.

Depth value scaling

Depth maps are stored as 16-bit unsigned integers $[0 - 65,535]$. In order to interpret them metrically, a linear mapping is applied so that the full range corresponds to a physical depth span. According to the dataset documentation for the SimCol3D challenge, the working depth interval for synthetic colonoscopy images is capped at 200 mm [42]; therefore, the full 16-bit range $[0-65,535]$ can be interpreted as mapping to $[0-200]$ mm in this project.

The conversion is performed using the following linear relationship:

$$\text{depth_metric} = \frac{200 \text{ mm}}{65535} \times \text{raw_value}. \quad (3.6)$$

This ensures that all depth values are converted into metric units (mm) before integration into the TSDF volume.

3.2.3 Pipeline

Each frame provides a dense depth map $D_i(u)$ where $u = (u_x, u_y)$ are pixel coordinates. Using the camera intrinsics matrix K (scaled to the resized image resolution), each pixel is back-projected into the camera coordinate system as:

$$\mathbf{x}_{\text{cam}}(u) = D_i(u) K^{-1} \begin{bmatrix} u_x \\ u_y \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{(u_x - c_x) D_i(u)}{f_x} \\ \frac{(u_y - c_y) D_i(u)}{f_y} \\ D_i(u) \end{bmatrix}. \quad (3.7)$$

The corresponding 3D point in world coordinates is obtained using the homogeneous transformation $T_i^{(\text{cam} \rightarrow \text{world})}$ derived from the pose files:

$$\mathbf{x}_{\text{world}}(u) = T_i^{(\text{cam} \rightarrow \text{world})} \begin{bmatrix} \mathbf{x}_{\text{cam}}(u) \\ 1 \end{bmatrix} = R_i \mathbf{x}_{\text{cam}}(u) + \mathbf{t}_i. \quad (3.8)$$

This transformation positions all measurements from different frames into a unified coordinate system. The inverse transform $T_i^{(\text{world} \rightarrow \text{cam})} = T_i^{-1}$ is used during integration, since the TSDF integrator expects the camera pose that maps world coordinates into the current camera frame.

Voxel updating

The TSDF volume maintains two fields per voxel \mathbf{v} :

- the truncated signed distance value $\phi(\mathbf{v})$, and
- an integration weight $w(\mathbf{v})$.

When processing frame i , the voxel center \mathbf{v} is transformed to the camera frame:

$$\mathbf{v}_{\text{cam}} = T_i^{(\text{world} \rightarrow \text{cam})} \mathbf{v} = \begin{bmatrix} x_c \\ y_c \\ z_c \\ 1 \end{bmatrix}.$$

If $z_c \leq 0$, the voxel lies behind the camera and is skipped. Otherwise, it is projected into the image plane as:

$$u_x = \frac{f_x x_c}{z_c} + c_x, \quad u_y = \frac{f_y y_c}{z_c} + c_y.$$

If the projected coordinates are inside the image, the algorithm retrieves the observed depth $z_{\text{surf}} = D_i(u_x, u_y)$. The signed distance between the voxel and the measured surface along the camera ray is then:

$$d_{\text{raw}} = z_{\text{surf}} - z_c.$$

A positive value means the voxel lies in front of the measured surface (free space), while a negative value means it lies behind the surface (inside the object). Since noisy or distant measurements could distort the surface, the distance is truncated to a limited range:

$$d_{\text{new}} = \begin{cases} -\tau & \text{if } d_{\text{raw}} < -\tau, \\ d_{\text{raw}} & \text{if } |d_{\text{raw}}| \leq \tau, \\ +\tau & \text{if } d_{\text{raw}} > +\tau, \end{cases} \quad (3.9)$$

where the truncation threshold τ defines the thickness of the region around the surface that contributes to updates (typically $\tau = 2$ mm). Each observation is merged into the TSDF via a weighted running average:

$$\phi^{\text{new}}(\mathbf{v}) = \frac{w(\mathbf{v}) \phi(\mathbf{v}) + w_{\text{obs}} d_{\text{new}}}{w(\mathbf{v}) + w_{\text{obs}}}, \quad w^{\text{new}}(\mathbf{v}) = \min(w(\mathbf{v}) + w_{\text{obs}}, w_{\text{max}}). \quad (3.10)$$

This simple formulation gives greater importance to regions that have been observed multiple times and naturally averages out measurement noise. Voxels that never fall within the truncation band retain their previous values (typically large positive ϕ), remaining unmodified. In this implementation, the TSDF volume is defined on a scalable voxel grid, where each voxel represents a cube of edge length

$$\text{voxel_length} = 0.5 \text{ mm}.$$

This value controls the spatial resolution of the reconstruction: smaller voxels capture finer detail of the mucosal folds but require substantially more memory and computation. The truncation band, which specifies the thickness of the region around the observed surface that receives updates, is set to

$$\tau = 2 \text{ mm},$$

corresponding to roughly four times the voxel size—an empirically stable ratio for preserving surface continuity without oversmoothing.

3.2.4 Outputs

After all frames have been integrated, the TSDF field $\phi(\mathbf{v})$ implicitly contains the colon surface as its zero-level isosurface:

$$S = \{\mathbf{x} \in \mathbb{R}^3 \mid \phi(\mathbf{x}) = 0\}. \quad (3.11)$$

To obtain an explicit 3D model, this surface is extracted using the **Marching Cubes** algorithm, which triangulates the isosurface into a watertight mesh. The resulting mesh inherits vertex normals from the gradient of the TSDF field:

$$\mathbf{n}(\mathbf{x}) = \frac{\nabla\phi(\mathbf{x})}{\|\nabla\phi(\mathbf{x})\|}. \quad (3.12)$$

Alternatively, the same zero-crossing region can be sampled to produce a **dense point cloud**. Both outputs are metrically consistent because all depth maps, poses, and intrinsics share the same physical scale.

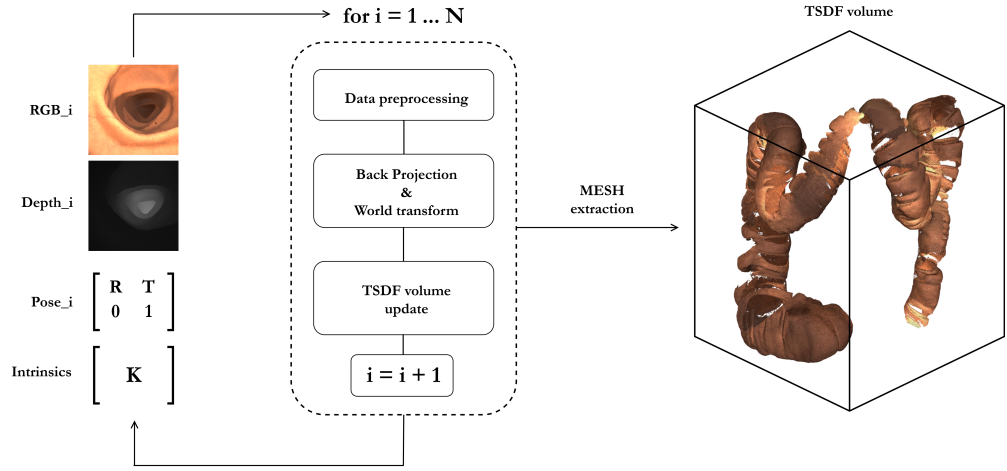


Figure 3.3: TSDF 3D reconstruction pipeline

3.3 Missing regions analysis

This second stage takes the open TSDF mesh (mesh.ply) and (i) repairs/closes it to obtain a watertight reference surface, and (ii) quantifies missing regions by comparing the repaired surface to the original open mesh. Conceptually, the original mesh encodes what was actually observed (with holes), while the repaired/closed mesh plays the role of a plausible completion of the lumen. Their difference localizes unobserved mucosa and supports per-region statistics and trajectory-based hotspot analysis.

3.3.1 Inputs

- Input mesh: a triangle mesh in PLY format produced by TSDF fusion.
- Trajectory: 6 DoF camera pose.

3.3.2 Mesh closing

The raw TSDF mesh typically contains small holes and irregular fragments that must be repaired before quantitative analysis. First, small and isolated surface patches are removed by clustering connected triangles and discarding single-triangle clusters; unreferenced vertices are dropped afterward. This prevents spurious slivers from biasing subsequent area and distance calculations.

If vertex normals are absent, they are computed on the original mesh so that a consistent oriented normal field is available. The resulting surface is then densified into an oriented point cloud via Poisson–disk sampling (here approximately 200 k points), which provides a quasi-uniform distribution of samples over the observed colon surface. This point set, together with the associated normals, is fed to Poisson Surface Reconstruction with a fixed octree depth (e.g. depth = 8), yielding a watertight triangle mesh and per-vertex density values. Higher reconstruction depths increase geometric detail at the cost of memory and computation time.

To avoid artificial “closures” far outside the actually observed region, the reconstructed Poisson mesh is cropped to the axis-aligned bounding box of the original TSDF mesh. This step preserves the interior sealing of luminal defects while trimming away exterior fills and large background components that are not supported by the original data.

Given oriented points $\{(\mathbf{p}_k, \mathbf{n}_k)\}$, Poisson reconstruction seeks an indicator function $\chi : \mathbb{R}^3 \rightarrow [0,1]$ whose gradient best matches the input normal field in a least-squares sense. A classical formulation solves

$$\nabla \cdot (\sigma \nabla \chi) = \nabla \cdot \mathbf{V}, \quad (3.13)$$

where \mathbf{V} is a smoothed vector field interpolating the normals (and σ a spatial weighting), then extracts the isosurface $\{\chi = \frac{1}{2}\}$ as a watertight mesh. Intuitively, χ is high “inside” the object and low outside, and the Poisson PDE integrates the normal constraints into a globally consistent, hole-free surface.

3.3.3 Missing region identification and quantification

To estimate which areas were not observed in the original reconstruction, the code compares the closed (Poisson) mesh to the original open mesh:

Reference Sampling

A set of 200,000 points is uniformly sampled from the original (open) mesh to form the evidence set $\mathcal{P}_{\text{orig}}$. This dense, approximately uniform sampling of the observed surface provides a geometric reference against which the closed mesh is evaluated, while keeping the computational cost of nearest-neighbor queries manageable.

Distance Field on the Repaired Surface

To assess which portions of the Poisson-closed mesh are not supported by the original evidence, the algorithm computes a distance field defined over the vertices of the repaired surface. For each vertex \mathbf{v} of the closed mesh, a 1-NN query is performed against the evidence set $\mathcal{P}_{\text{orig}}$ sampled from the open mesh. This is implemented using a `NearestNeighbors` structure in scikit-learn, which builds a KD-tree accelerator for efficient querying over the 200,000 sampled points.

$$d(\mathbf{v}) = \min_{\mathbf{x} \in \mathcal{P}_{\text{orig}}} \|\mathbf{v} - \mathbf{x}\|_2. \quad (3.14)$$

Vertices with small values of $d(\mathbf{v})$ lie close to regions that were already observed in the TSDF reconstruction, whereas large values correspond to areas on the Poisson mesh that lack direct support in the original data and are thus potential candidates for missing regions.

The resulting scalar field captures, with high spatial resolution, the geometric disagreement between the observed surface and the watertight completion. This field is later refined through several filtering stages to avoid false positives introduced by thin structures, normal noise, and meshing artifacts.

Adaptive Thresholding

Identifying a global threshold that separates supported from unsupported surface regions is nontrivial, as mesh resolution and local geometric density vary significantly

across the colon anatomy. The algorithm therefore employs an adaptive cutoff τ_d , determined as the minimum of several characteristic length scales:

$$\tau_d = \min(d_{\text{user}}, 5 \bar{\ell}, 8 \text{med}(\ell), 3.0 \text{ mm}), \quad (3.15)$$

where $\bar{\ell}$ and $\text{med}(\ell)$ denote the mean and median edge lengths of the repaired mesh, estimated over up to 2,000 randomly selected faces. This sampling strategy ensures a low computational cost while providing a statistically robust estimate of the mesh scale.

The user-defined component d_{user} (set to 1.0 mm in the main pipeline) acts as a hard cap to avoid overly permissive thresholds in regions with coarse triangulation. The additional bound at 3.0 mm limits the threshold even under extreme geometric variations.

By combining mesh-derived and user-defined scales, τ_d adapts to local surface resolution while remaining clinically conservative, preventing overestimation of missing regions in areas with naturally sparse sampling.

Local-Density Veto

Thresholding alone may falsely classify vertices lying near thin anatomical structures or along seams created by Poisson reconstruction. To mitigate this, each candidate vertex \mathbf{v} with $d(\mathbf{v}) > \tau_d$ undergoes a density check against the evidence set $\mathcal{P}_{\text{orig}}$.

Specifically, the algorithm counts the number of original points lying within a radius $2\tau_d$ of \mathbf{v} . If at least five points are found, the vertex is discarded as a false positive, since the local evidence indicates that the region was actually observed, despite a comparatively large nearest-neighbor distance.

This veto mechanism suppresses artifacts originating from small offsets between the Poisson surface and the original geometry—such as along narrow folds, self-contact areas, or slightly inflated patches—where the missing-region hypothesis would otherwise be triggered.

Controlled Boundary Growth

After the initial thresholding and density filtering, the remaining candidate vertices tend to form sparse and fragmented patches. To recover spatially coherent missing regions while avoiding uncontrolled expansion, a constrained region-growing scheme is applied.

For each vertex marked as missing, its one-ring neighbors are inspected. A neighboring vertex \mathbf{u} is added to the missing mask only if its distance satisfies

$$d(\mathbf{u}) > 0.8 \tau_d,$$

which permits moderate relaxation of the primary threshold while maintaining a strict geometric criterion. This controlled dilation step ensures that irregular or noisy boundaries are smoothed and that missing regions become contiguous, without leaking into well-supported areas of the mesh.

Region Assembly

Once a stable vertex mask is obtained, the algorithm constructs the missing regions mesh by selecting triangles according to the state of their vertices. A triangle is immediately included if all three of its vertices are marked as missing. Border triangles with exactly two missing vertices are considered separately: they are included only if the remaining vertex satisfies $d > 0.5 \tau_d$, ensuring that boundary faces are incorporated only when they are likely to be genuine extensions of the missing region rather than noisy transitions.

This selective assembly strategy produces well-formed, topologically consistent patches that reflect unobserved zones of the colon while minimizing accidental inclusion of support-rich boundary triangles.

Significant-Region Filtering

The assembled missing regions may still contain small, noisy components such as isolated slivers, thin strips at mesh seams, or artifacts generated by Poisson reconstruction. To ensure clinical and geometric relevance, connected components are extracted and retained only if they satisfy two conditions:

$$\text{triangle count} \geq 20, \quad \text{area} \geq 5.0 \text{ mm}^2.$$

Components below these thresholds are discarded as negligible. The area criterion is intentionally conservative: even the smallest clinically relevant polyps (diameter $\approx 5 \text{ mm}$) would project to a surface area well above 20 mm^2 , meaning that sub- 5 mm^2 regions are exceedingly unlikely to represent meaningful mucosal gaps. Instead, such small fragments typically correspond to discretization noise, imperfect Poisson blending, or boundary effects.

The result of this filtering stage is a set of anatomically and clinically meaningful missing regions suitable for downstream quantitative and spatial analysis.

Figure 3.4 summarizes the entire missing-region extraction pipeline presented above.

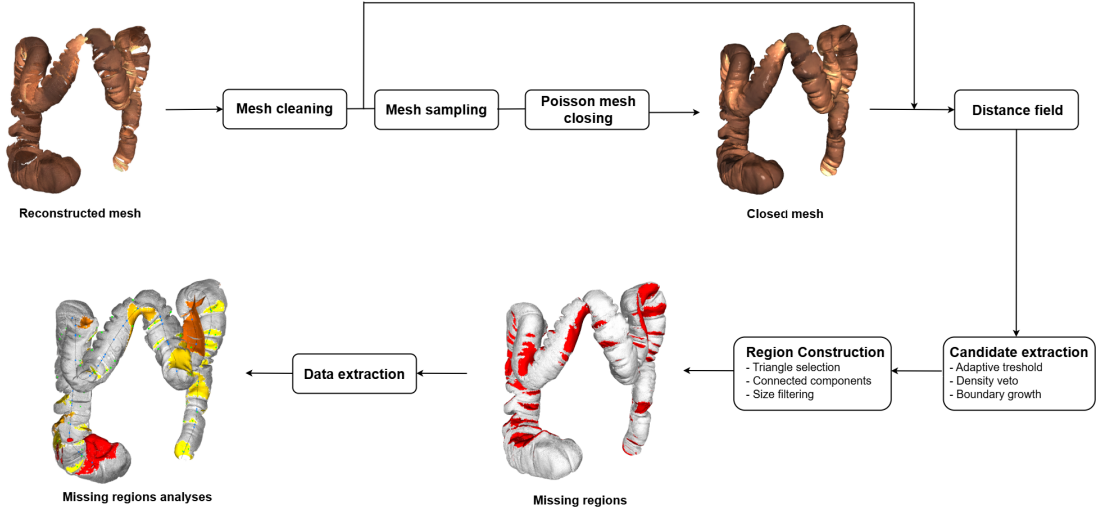


Figure 3.4: Missing region identification pipeline.

Global metrics

These metrics summarize the overall quality of the reconstruction:

- Closed mesh and Missing regions mesh in PLY format.
- Missing Percentage : the ratio of missing vertices (after filtering) to total vertices in the repaired mesh, expressed as a percentage:

$$\%missing = \frac{\#V_{missing \text{ (filtered)}}}{\#V_{repaired}} \times 100. \quad (3.16)$$

- **Counts:** Missing vertices/triangles - number of significant regions - regions removed by filtering.

Single region metrics

For each significant missing region k , the following per-region metrics are computed and reported:

- Area in number of vertices and mm^2 .
- 3D coordinates of the centroid (barycenter), computed as:

$$\mathbf{c}_k = \frac{1}{|\mathcal{V}_k|} \sum_{\mathbf{v} \in \mathcal{V}_k} \mathbf{v}, \quad (3.17)$$

Let the endoscopic trajectory be represented by the ordered sequence of N camera positions

$$\mathcal{T} = \{\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_N\}, \quad \mathbf{p}_i \in \mathbb{R}^3.$$

In the implementation, the raw endoscopic trajectory (loaded from `SavedPosition.txt`) is first converted from centimetres to millimetres to match the mesh scale, and then rigidly translated so that its centroid coincides with the centroid of the reconstructed colon. This alignment ensures that both the distances d_k and the longitudinal percentages λ_k are computed in a consistent anatomical frame.

For each centroid \mathbf{c}_k , the corresponding point on the trajectory, denoted as \mathbf{p}_k^* , is identified as the one minimizing the Euclidean distance to the centroid:

$$\mathbf{p}_k^* = \arg \min_{\mathbf{p}_i \in \mathcal{T}} \|\mathbf{p}_i - \mathbf{c}_k\|_2.$$

The minimum distance value

$$d_k = \min_{\mathbf{p}_i \in \mathcal{T}} \|\mathbf{p}_i - \mathbf{c}_k\|_2$$

quantifies how far the missing region is from the trajectory path.

Once the closest trajectory index i_k^* associated with \mathbf{p}_k^* is known, its relative position along the trajectory is computed as a percentage of the total path length:

$$\lambda_k = \frac{L_{i_k^*}}{L_{\text{tot}}} \times 100,$$

where $L_{i_k^*}$ is the cumulative distance from the trajectory start up to point i_k^* , and L_{tot} is the total trajectory length:

$$L_i = \sum_{j=1}^{i-1} \|\mathbf{p}_{j+1} - \mathbf{p}_j\|_2, \quad L_{\text{tot}} = L_N.$$

The value λ_k therefore expresses the longitudinal position of each missing region along the endoscope trajectory, indicating whether it is located near the entrance or deeper within the reconstructed colon. This information provides valuable context to the endoscopist, allowing an immediate understanding of the depth and approximate location of each unobserved region for subsequent inspection.

Figure 3.5 illustrates the geometric relationship between a missing region's centroid \mathbf{c}_k and its nearest trajectory point \mathbf{p}_k^* . The distance d_k quantifies how far the endoscope must be repositioned to potentially cover the unobserved area, while the trajectory percentage λ_k indicates at what depth along the procedure this repositioning should occur. Together, these metrics enable targeted reinsertion strategies for improved mucosal coverage.

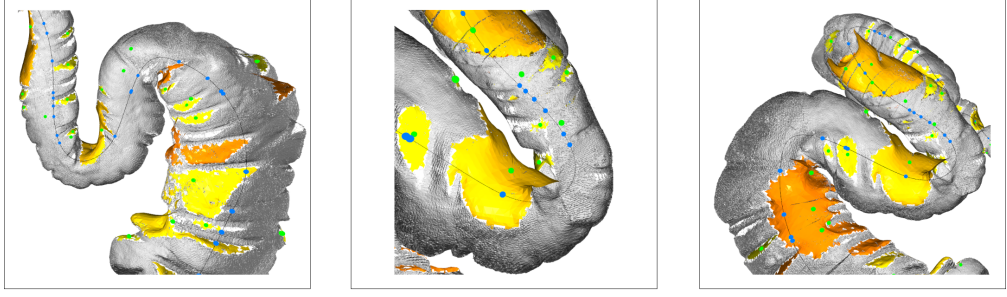


Figure 3.5: Example visualization of a reconstructed mesh with identified missing regions (yellow/orange surfaces), their corresponding centroids (green spheres), and the nearest trajectory points to each centroid (blue spheres). The geometric relationship between each missing region’s centroid \mathbf{c}_k and its closest trajectory point \mathbf{p}_k^* provides actionable spatial feedback for targeted re-examination during colonoscopy.

3.3.4 Missing regions distribution

To obtain a global overview of how unobserved regions are distributed along the colon anatomy, a dedicated analysis was performed. The goal of this procedure was to highlight which areas of the colon are most frequently missed across multiple endoscopic reconstructions of the same anatomy.

The method operates by aggregating the results obtained from all the fifteen sequences (**S1–S15**) of the **SyntheticColon_I** model of the **SimCol3D** dataset. For each sequence, the corresponding mesh of missing regions—computed as the difference between the reconstructed and Poisson-closed meshes—is first loaded and sampled into a dense set of 3D points. Each point represents a location belonging to a missing region.

All sampled points from the different sequences are then quantized into a common 3D voxel grid using a voxel size of $\text{VOXEL_SIZE} = 0.7 \text{ mm}$. The coordinates of each voxel are defined as:

$$\mathbf{v}_i = \lfloor \mathbf{p}_i / s \rfloor, \quad s = \text{VOXEL_SIZE},$$

where \mathbf{p}_i is a sampled point and $\lfloor \cdot \rfloor$ denotes the floor operator. For each voxel index \mathbf{v}_i , the number of times it is occupied across all sequences is counted, yielding an overlap count n_i . The set of unique voxels and their corresponding frequencies are then computed as:

$$\mathcal{V} = \{\mathbf{v}_i\}_{i=1}^M, \quad n_i = \text{count}\left(\mathbf{v}_i \text{ occurs in } \bigcup_{k=1}^{15} \text{seq}_k\right).$$

Finally, each voxel center is mapped back into the world coordinate frame:

$$\mathbf{c}_i = \mathbf{s} \cdot \mathbf{v}_i,$$

and visualized as a 3D colored point cloud where the color encodes the frequency n_i , indicating how many sequences exhibit a missing region in that spatial location. The resulting global heatmap thus provides an intuitive and quantitative representation of the anatomical zones most frequently unobserved during endoscopic exploration.

This aggregated visualization makes it possible to identify recurrently unobserved segments of the colon, typically corresponding to complex folds or areas poorly reached by the camera field of view. Such information is valuable both for evaluating the completeness of automated reconstructions and for providing feedback to endoscopists regarding anatomical regions that are more likely to remain unexamined.

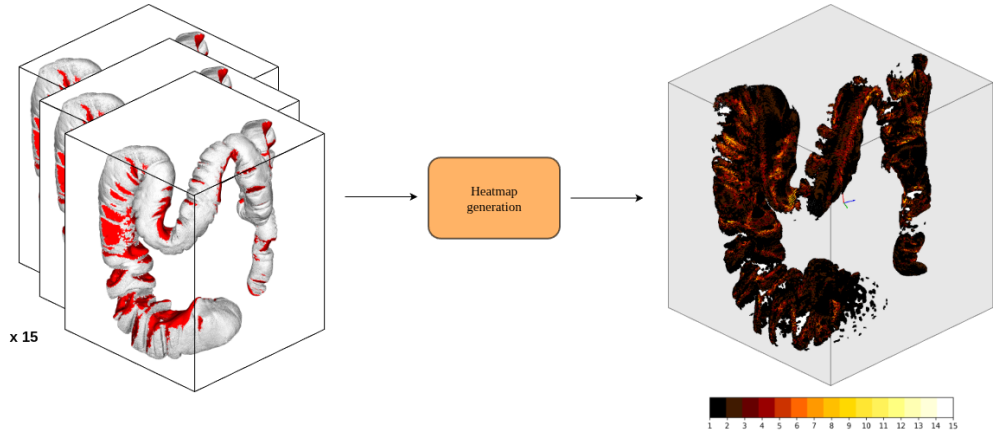


Figure 3.6: The heatmap pipeline for missing regions distribution analysis. The fifteen missing-region meshes (S1–S15) are spatially overlapped and quantized into a common voxel grid. Each voxel accumulates a count representing how many sequences exhibit a missing region at that location. The resulting point cloud is color-coded according to this frequency: lighter colors indicate areas that are repeatedly unobserved across multiple sequences, revealing anatomical zones prone to incomplete coverage.

3.4 Depth estimation

One of the fundamental components of the 3D reconstruction pipeline is the depth estimation module, which predicts per-pixel depth maps from monocular RGB endoscopic images. Accurate depth estimation is crucial for reconstructing the colon’s intricate geometry and ensuring reliable spatial reasoning during navigation. In this work, several state-of-the-art deep learning models for monocular depth estimation were evaluated.

3.4.1 Depth-Pro

Depth Pro [22] is a monocular metric depth model from Apple that takes a single RGB image (optionally with a known focal length) and returns a high-resolution depth map in meters plus, when intrinsics are missing, an estimated focal length in pixels; the official API exposes both keys depth and focal length directly. Architecturally, it uses a multi-scale ViT: a global image encoder supplies scene context while a weight-shared patch encoder processes overlapping multi-scale crops; features are then merged and refined by a light DPT-style decoder to produce a sharp 2.25-MP map, and a small auxiliary head regresses focal length from intermediate features—allowing absolute-scale predictions even without EXIF intrinsics. The model is trained with a two-stage curriculum: first, mixed real + synthetic data to learn robust metric depth (combining MAE on metric datasets with scale-invariant terms where appropriate), then synthetic-only fine-tuning that adds gradient/Laplacian losses to sharpen boundaries.

Model Architecture

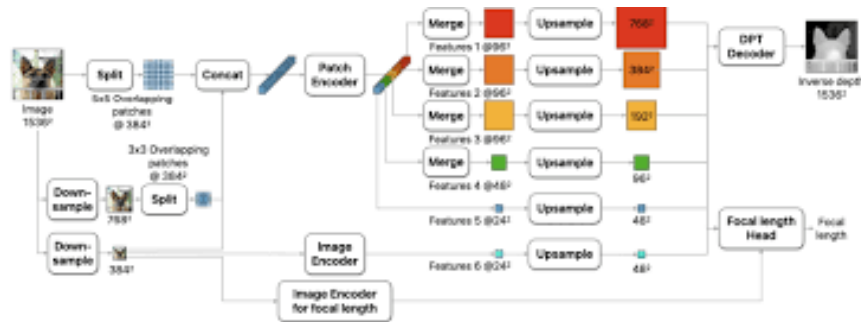


Figure 3.7: Depth-Pro pipeline

Application on endoscopic images

In medicine, Depth Pro has already been tested on endoscopic imagery. The “Zero-shot Monocular Metric Depth for Endoscopic Images” [43] benchmark evaluates state-of-the-art depth models—including Depth Pro—on real clinical endoscopy and introduces EndoSynth for finetuning, reporting zero-shot and adapted performance; and Xu et al.[23] explicitly evaluate Depth Pro by transfer on endoscopy datasets within a self-supervised depth-and-pose framework.

3.4.2 Depth-Anything

Depth Anything [19] takes a single RGB image and outputs a dense depth map, exposed via simple CLI/Python APIs. The architecture follows a DPT-style encoder-decoder with a DINOv2 vision backbone and multi-scale features, designed for robust dense prediction across diverse scenes. Its training hinges on scale: a data engine assembles 62M unlabeled images plus 1.5M labeled samples for pretraining with pseudo/auxiliary supervision, then a compact metric fine-tuning stage (e.g., on NYUv2 and KITTI) calibrates absolute scale; official materials and the model card detail this recipe and design. In practice the project provides V1 (CVPR 2024) and the follow-up V2 with improved accuracy and speed, but the core idea—massive data with a simple, strong architecture—remains.

Model Architecture

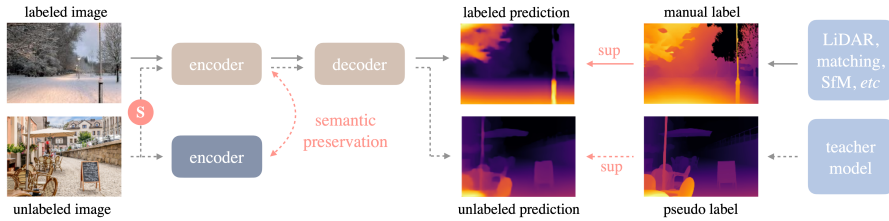


Figure 3.8: Depth-Anything pipeline

Application on endoscopic images

Depth anything has found applications in the medical environment. A colonoscopy-specific example is ColonCrafter [24], whose evaluation protocol directly runs Depth-Anything (V1/V2) checkpoints to produce depth maps on the C3VD benchmark and reports their metrics alongside endoscopy-specific models—explicitly noting the domain gap and the need for adaptation in this setting.

3.4.3 Video-Depth-Anything

Video Depth Anything (VDA) [21] takes a monocular video (RGB frames) as input and outputs a temporally consistent dense depth map per frame (non-metric by default), built on Depth Anything V2 with a new spatio-temporal head: global image features from DA-V2 are passed to a temporal self-attention module whose outputs are decoded into depth, while a temporal-gradient matching loss enforces frame-to-frame smoothness; at inference, a key-frame strategy lets it process very long videos efficiently and even reach real-time in its small variant. The official repo/project page provides models at multiple scales and a Python/CLI for batch video inference. Training follows DA-V2’s recipe—joint learning on large unlabeled image corpora plus video-depth data, then optimization with the temporal loss—yielding state-of-the-art zero-shot consistency on public video benchmarks.

Model Architecture

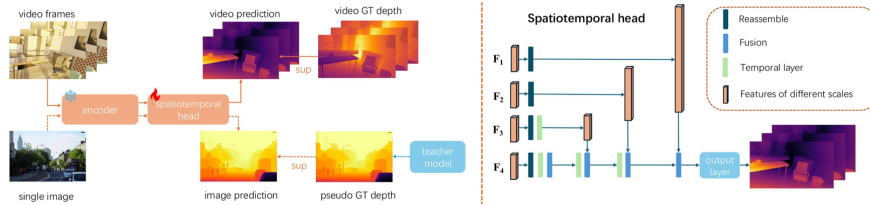


Figure 3.9: Video-Depth-Anything pipeline

Application on endoscopic images

In medicine, VDA has already been adapted to endoscopy: EndoDAV [44] fine-tunes a VDA model with self-supervised, parameter-efficient updates and a projection/alignment strategy to obtain spatiotemporally consistent endoscopic depth, demonstrating the practicality of transferring VDA to surgical/diagnostic videos.

3.4.4 SUMNet

Overview and inputs. The proposed system³ follows the SUMNet approach developed by the **KLIV team** for the SimCol-to-3D 2022 challenge. Their method applies a fully-convolutional neural network to estimate dense colon depth maps from frame buffers while preserving fine structural details and avoiding the loss

³<https://github.com/SistaRaviteja/Colonoscopy-Depth-Estimation>

of critical information. The input data consist of synthetic colonoscopy frames named `FrameBuffer_*.png`. During training, each RGB frame is paired with its corresponding dense ground-truth depth map `Depth_*.png`. Frames are converted from RGBA to RGB, resized to 448×448 , converted to tensors, and normalized using the training set statistics (mean and standard deviation computed in a preliminary pass). Depth maps are loaded in grayscale, scaled to the same spatial resolution, and normalized to $[0,1]$. The file lists for the train/validation/test splits are specified in `train_file.txt`, `val_file.txt`, and `test_file.txt`.

Network architecture. The model is a fully-convolutional encoder–decoder that (i) transfers max-pooling indices from the encoder to the decoder’s `MaxUnpool2d` layers at the same depth, (ii) uses a VGG-11 backbone (pretrained on ImageNet) as encoder, and (iii) concatenates encoder activations with decoder feature maps at corresponding spatial scales (“skip” concatenations). Each encoder block ends with a `MaxPool2d`, and the decoder mirrors this structure with `MaxUnpool2d(2,2)` guided by those indices. The unpooled features are concatenated with the corresponding encoder activations and refined using lightweight convolutional blocks (`Conv + BatchNorm + ReLU`). A final 1×1 convolution produces a single-channel depth map, and a sigmoid activation constrains the output to the normalized range $[0,1]$. Preserving pooling indices—a core idea borrowed from SegNet—allows for the recovery of spatial details during upsampling, which is essential for the thin, high-frequency anatomical structures encountered in endoluminal imagery.

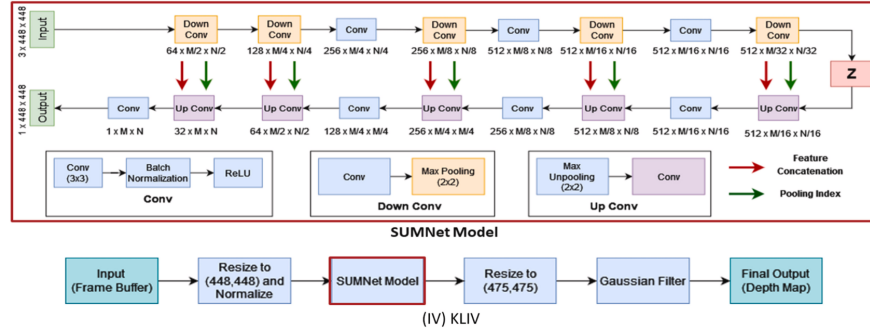


Figure 3.10: SUMNet architecture

Training pipeline and hyper-parameters. The model is implemented in PyTorch and trained for **50 epochs**. Training uses Adam optimization with an initial learning rate of 1×10^{-3} and an exponential learning rate scheduler with decay factor $\gamma = 0.98$. Mixed-precision training is enabled via `torch.cuda.amp` and gradient scaling. A batch size of **16** is used, with dataloaders configured as `shuffle=True`, `num_workers=8`, `pin_memory=True`, and `drop_last=True`. Model

checkpoints—including network weights, optimizer state, and scheduler parameters—are saved every 5 epochs to the directory `sumnet_model/`. The network is trained using different loss functions to evaluate their influence on convergence and final accuracy: **Mean Squared Error (MSE)**, **Mean Absolute Error (MAE)**, **scale-invariant log loss** and **BerHu loss**. Among these, the model trained with MSE produced the most reliable and accurate results across the validation set. To further mitigate aliasing artifacts in the predicted depth maps, a post-processing step applies a **Gaussian Blur low-pass filter** with a kernel size of 7×7 to smooth residual high-frequency noise. The dataset used for training corresponds to the **SyntheticColon_II** anatomy of the **SimCol3D** dataset, specifically using the sequences **B1–B15** as the training subset.

Inference and output format. During inference, the input frames are resized to 448×448 , normalized using the dataset statistics, and passed through the trained network to produce depth maps. The predicted outputs are then resized to 475×475 (the SimCol3D evaluation format) and saved as `.npy` arrays in `float16` precision. According to the challenge convention, the output range is $[0,1]$, where a normalized depth value of 1.0 corresponds to a physical distance of 20 cm. The quantitative evaluation metrics include the **L1 error**, **relative error**, and **root-mean-square error (RMSE)** between the predicted and ground-truth depths, as defined in the SimCol3D benchmark protocol.

3.4.5 Depth estimation summary

Using the previously described methods, depth maps were estimated for the sequences **S1–S15** of the **SyntheticColon_I** anatomy from the **SimCol3D** dataset. Although the internal architectures and training strategies differ, their outputs were uniformly post-processed and formatted to ensure compatibility with the downstream 3D reconstruction pipeline. In particular, each predicted depth map was resized, normalized, and converted to the same image format and spatial resolution as those of the **SimCol3D** dataset. This harmonization step was essential because the subsequent reconstruction system—based on the **TSDF (Truncated Signed Distance Function)** pipeline—was originally implemented and optimized to operate with the image and depth formats provided by SimCol3D. Therefore, standardizing the predicted outputs allowed all the models to interface seamlessly with the TSDF-based reconstruction module, enabling a fair and consistent comparison of the 3D reconstruction results obtained from different depth-estimation strategies.

The performance of the depth estimation methods have been quantitatively evaluated using the GT depth from the SimCol3D dataset as reference.

The metrics obtained shows that the SUMNet model outperforms the others in terms of L1 error, relative error, and RMSE, indicating its superior accuracy in predicting depth maps for colonoscopy images.

3.4.6 Depth processing

Since the depth maps obtained from the Sumnet model showed superior accuracy compared to other methods, all subsequent experiments were carried out using these predictions. However, visual inspection revealed noticeable granularity and temporal inconsistency between consecutive frames, which negatively affected the quality of the resulting 3D reconstructions. To mitigate these issues, an additional post-processing step was introduced, aimed at filtering and stabilizing the depth maps to improve both spatial smoothness and temporal coherence.

The system takes applies a **spatio-temporal filter** to all the depth maps.

Spatial smoothing is implemented as a **Gaussian blur** with kernel size `KSIZE` (odd), i.e.:

$$S_t = G_\sigma * D_t$$

where D_t is the raw depth map at time t , G_σ is a 2-D Gaussian kernel, and $*$ denotes convolution. More explicitly, the spatial filtering operation can be written as:

$$S_t(x, y) = \sum_{i=-k}^k \sum_{j=-k}^k G_\sigma(i, j) D_t(x - i, y - j)$$

where

$$G_\sigma(i, j) = \frac{1}{2\pi\sigma^2} e^{-\frac{i^2+j^2}{2\sigma^2}}$$

is the Gaussian kernel of standard deviation σ , and k defines the kernel radius.

Temporal consistency is enforced with an **exponential moving average**:

$$\widehat{D}_t = \alpha \widehat{D}_{t-1} + (1 - \alpha) S_t$$

with $\alpha = \text{TEMPORAL_ALPHA}$; for the first frame, $\widehat{D}_0 = S_0$.

Frames are processed in `float32` to preserve dynamic range; if an image is multi-channel, only channel 0 is used. Before saving, each filtered frame is **clipped and cast back** to its original data type (e.g., `uint16`) to ensure compatibility.

3.5 Pose estimation

In this section I present a pose estimation model proposed by Rau et al. [26] that predicts the 6-DoF camera pose for each frame in the endoscopic video sequence. Accurate pose estimation is essential for aligning depth maps into a common coordinate system during 3D reconstruction.

The network takes two consecutive RGB colonoscopy frames and outputs the *relative* 6-DoF pose of frame 2 with reference to frame 1: a 3-vector translation and a 3-vector log-quaternion for rotation (together a 6-D output). The authors define the camera-pair projection so that $+z$ is forward, converting Unity’s left-handed poses to a right-handed convention for training and evaluation.

Model Architecture

The model follows a bimodal design that explicitly handles the two dominant motion patterns along the colon (forward *insertion* vs. backward *withdrawal*). Each of the two input frames (I_t, I_{t+k}) is processed by a shared ResNet-18 encoder to extract convolutional feature maps. A non-learned *correlation layer* computes dense correlations between the two feature maps; this correlation volume is particularly informative for the direction classification task. In parallel, frame features are concatenated channel-wise and fed to the pose regressor.

Two heads operate on these representations:

- **Classification head (Class Net).** A lightweight MLP (three layers with strong dropout) predicts the probability that the pair (I_t, I_{t+k}) belongs to the *insertion* or *withdrawal* motion bin. This head uses the correlation volume to robustly disambiguate forward vs. backward motion.
- **Pose regression head (Pose Net).** A fully convolutional stack (four conv blocks with ReLU) outputs *two* residual 6-DoF poses, one for each motion bin. Let $\mathbf{b}_1, \mathbf{b}_2 \in \mathbb{R}^6$ denote the fixed bin centers (translation along $\pm z$; rotation in log-quaternion), and let $\Delta \mathbf{w}_1, \Delta \mathbf{w}_2 \in \mathbb{R}^6$ be the residuals regressed by the head. With class probabilities $\mathbf{p} = [p_1, p_2]$, the final relative pose is obtained as a probability-weighted mixture:

$$\hat{\Omega}_{t \rightarrow t+k} = p_1(\mathbf{b}_1 + \Delta \mathbf{w}_1) + p_2(\mathbf{b}_2 + \Delta \mathbf{w}_2).$$

This bimodal “mixture-of-experts” formulation forces the network to explicitly model the empirically bimodal displacement distribution along the lumen axis. Rotations are represented as *log-quaternions*, a minimal 3D parameterization obtained by applying the logarithmic map to unit quaternions; the 6-D pose vector is thus $[\mathbf{t} \mid \log \mathbf{q}]$.

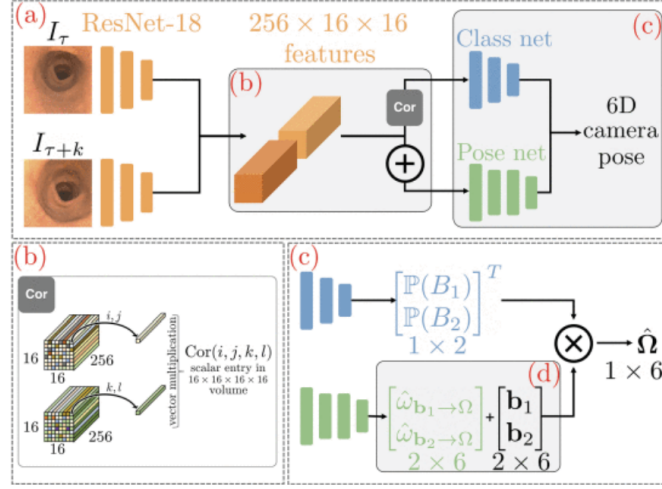


Figure 3.11: Bimodal camera pose prediction architecture (Class Net + Pose Net with correlation volume and residuals around class means).

Training overview

The model is trained in a fully supervised manner using the SimCol3D synthetic colonoscopy dataset. The split used in this work is: **S1, S2, S3, S7, S8, S11** for training; **S12, S13** for validation; and **S4, S5, S9, S10, S14, S15** for testing/inference. Training samples are formed as two-frame pairs (I_t, I_{t+k}) at a fixed stride k ; native images are resized/center-cropped to the working resolution and normalized. For each pair, the ground-truth *relative* 6-DoF pose $(\mathbf{t}, \log \mathbf{q})$ is available from the simulator.

The Class Net is optimized with cross-entropy to distinguish insertion vs. withdrawal and typically reaches very high validation accuracy early in training, thanks to the correlation cue. The Pose Net is trained with a learned-weighted combination of translation and rotation errors following Kendall’s uncertainty weighting:

$$\mathcal{L}_{\text{pose}} = \|\mathbf{t} - \hat{\mathbf{t}}\|_2 e^{-\beta} + \beta + \|\log \mathbf{q} - \log \hat{\mathbf{q}}\|_2 e^{-\gamma} + \gamma,$$

where β, γ are scalar parameters learned jointly with the network to balance the two terms. The total loss combines pose and class terms:

$$\mathcal{L} = \mathcal{L}_{\text{pose}} + w_c \mathcal{L}_{\text{class}},$$

with w_c a small constant weight for the classification loss. Training uses the ResNet-18 backbone (shared weights for the two frames), the correlation layer for the Class Net, the 4-layer conv Pose Net, and strict data pairing to preserve temporal ordering.

Inference and trajectory composition

At test time, the network outputs the relative pose $\hat{\Omega}_{t \rightarrow t+k}$ for each consecutive pair. Absolute camera trajectories are recovered by sequential composition of the relative transforms along the sequence (left-multiplication in the chosen world frame). For quantitative evaluation, the predicted trajectory is *scale-aligned* to ground truth by a single scalar s that matches the translation magnitude (no rigid SE(3) alignment), preserving the direction conventions (insertion vs. withdrawal):

$$s = \frac{\sum_t \text{trans}(\mathbf{P}_t)^\top \text{trans}(\hat{\mathbf{P}}_t)}{\sum_t \text{trans}(\hat{\mathbf{P}}_t)^\top \text{trans}(\hat{\mathbf{P}}_t)},$$

where \mathbf{P}_t and $\hat{\mathbf{P}}_t$ are absolute GT and predicted poses, and $\text{trans}(\cdot)$ extracts the translation component. This avoids confounding rigid misalignment with genuine odometry errors, while accounting for potential global scale mismatch.

Evaluation metrics

Following the protocol in [26], performance is summarized with:

- **Relative Translation Error (RTE):** local step-wise translation error between consecutive poses (after scale alignment).
- **Absolute Translation Error (ATE):** global position drift accumulated after chaining the relative predictions into a full trajectory.
- **Rotation Error (ROT):** angular magnitude of the local rotation error (e.g., from the relative rotation matrix R_Δ via $\arccos((\text{trace}(R_\Delta) - 1)/2)$).

These metrics are computed over the test split and reported both per-sequence and averaged across sequences .

Implementation notes

All experiments adopt the right-handed convention with $+z$ forward in the camera-pair projection; Unity left-handed poses are consistently converted during data preparation. The bimodal design (bin centers $\mathbf{b}_1, \mathbf{b}_2$ along $\pm z$) and the correlation-guided Class Net are crucial to robustly handle the frequent direction reversals in colonoscopy while enabling the Pose Net to focus on residual offsets around plausible motion modes.

3.6 Hybrid Dataset acquisition

Since a significant portion of the preliminary work was carried out using simulated data and virtual images, the subsequent phase of the study aimed to transition toward a real-world experimental setting. The main objective of this stage was to design and acquire a benchmark dataset under realistic conditions, in order to validate and assess the performance of the reconstruction pipeline using real endoscopic data. To achieve this, an Olympus PCF-PH190I colonoscope was employed, connected to an Olympus EVIS EXERA III video processing unit. This configuration allowed the recording of high-quality colonoscopy videos on a silicone-based anatomical phantom, which accurately reproduces the geometry and optical properties of the human colon. In parallel with video acquisition, the spatial position and orientation of the endoscope tip were continuously monitored using the Aurora electromagnetic tracking system (Northern Digital Inc., NDI). This system provides high-precision 6-DoF pose measurements by exploiting low-frequency electromagnetic fields and miniature sensor coils embedded at the distal end of the instrument. The integration of these two data resulted in a synchronized multimodal dataset that associates each RGB frame with the corresponding camera pose. This dataset serves as a crucial foundation for evaluating the accuracy and robustness of the reconstruction algorithms, bridging the gap between simulated and real clinical environments.

3.6.1 Endoscope and video acquisition

The optical acquisition system consisted of an Olympus PCF-PH190I colonoscope connected to an EVIS EXERA III video processor (CV-190) and xenon light source.

The colonoscope provides HDTV imaging with a 140° field of view.

The distal end has a diameter of 9.7 mm, with a working length of 1680 mm and a 3.2 mm instrument channel. The tip allows 180° up/down and 160° left/right angulation, ensuring flexible maneuverability within the lumen.

The EVIS EXERA III platform delivers high-definition video output (1080i), advanced image enhancement, and stable xenon illumination (300 W) with automatic brightness control.

3.6.2 Phantom model

The experiments were conducted using two distinct anatomical phantoms to progressively validate the acquisition pipeline. Initially, a silicone colon segment was employed to establish and test the baseline data acquisition workflow. This phantom consisted of a tubular silicone structure replicating a portion of the colon anatomy, providing realistic surface texture and mechanical flexibility for endoscope

insertion. Once the acquisition protocol was validated, the experimental setup was transitioned to a KAGARU phantom, a commercially available, high-fidelity training model that reproduces the complete anatomical geometry and optical properties of the human colon with enhanced realism and structural complexity.

3.6.3 Aurora NDI

The Aurora electromagnetic tracking system from NDI ⁴ consists of the following main components:

Field Generator (FG)

Emits a low-intensity, time-varying electromagnetic field that defines the measurement volume. With the Planar 20-20 option the characterized cube mode is $500 \times 500 \times 500$ mm, suitable for tabletop/phantom setups; plug-and-play with the controller.

Sensor Interface Unit (SIU)

Hardware front-end that amplifies and digitizes the voltages coming from the miniature tracking sensors; available with 2/4/6/8 ports (each port: one 6-DoF tool or two 5-DoF tools).

System Control Unit (SCU)

Central module that drives the FG, aggregates SIU data, and computes 3D position and orientation, exposing poses to the host PC.

Sensors / probe coil

Miniaturized 5-DoF or 6-DoF coils that can be embedded at the distal tip of flexible instruments to act as localized tracking points. In this case it has been inserted inside the working channel of the colonoscope. The NDI AURORA electromagnetic tracking system creates a calibrated magnetic field within a controlled workspace of about $50 \times 50 \times 50$ cm, positioned in front of the field generator—essentially the operative area where the endoscope and phantom move. Within this volume, a miniature 6-DoF probe inserted through the endoscope’s instrument channel (close to the optical tip) continuously interacts with the magnetic field.

⁴<https://www.ndigital.com/electromagnetic-tracking-technology/aurora/>



Figure 3.12: Aurora components. From left to right: FG, SCU, SIU.

3.6.4 Pose computation

The interaction follows Faraday’s law of electromagnetic induction, according to which a time-varying magnetic flux $\Phi(t)$ through a coil generates a voltage $v(t)$:

$$v(t) = -N \frac{d\Phi(t)}{dt}, \quad \Phi(t) = \int_A \mathbf{B}(\mathbf{x}, t) \cdot \hat{\mathbf{n}} dA, \quad (3.18)$$

where N is the number of turns and $\mathbf{B}(\mathbf{x}, t)$ is the magnetic field at position \mathbf{x} . The field generator excites its three orthogonal coils in a known, time-multiplexed sequence, producing fields $\mathbf{b}_i(\mathbf{x})$. Each probe coil senses a distinct combination of these components, and the induced voltages depend on the probe’s position (field strength) and orientation (dot product $\mathbf{B} \cdot \hat{\mathbf{n}}$).

The analog voltages are routed to the Sensor Interface Unit (SIU) for amplification and digitization, then sent to the System Control Unit (SCU), which performs synchronous demodulation synchronized with the generator’s excitation. This process isolates the contribution of each generator axis and yields a coupling matrix

$$C_{\text{meas}} = [c_{ji}] \in \mathbb{R}^{3 \times 3},$$

describing how the probe’s coil axes interact with the generator’s field axes. The SCU compares this matrix with the modeled field $B(\mathbf{T}) = [\mathbf{b}_1(\mathbf{T}) \mathbf{b}_2(\mathbf{T}) \mathbf{b}_3(\mathbf{T})]$ and estimates the probe’s rotation \mathbf{R} and translation \mathbf{T} by minimizing

$$\min_{\mathbf{R}, \mathbf{T}} \| C_{\text{meas}} - \mathbf{R}^\top B(\mathbf{T}) \|_F^2. \quad (3.19)$$

The optimization alternates between a closed-form orientation update (via SVD or quaternion alignment) and a position refinement (via Gauss–Newton on the field map). The result is a 6-DoF pose $T_{\mathcal{W} \leftarrow \mathcal{S}}$ expressed as a 4×4 homogeneous matrix,

$$\begin{bmatrix} \mathbf{R} & \mathbf{T} \\ \mathbf{0}^\top & 1 \end{bmatrix},$$

computed at about 40 Hz and transmitted to the acquisition computer, where it is displayed in real time and recorded as the ground-truth pose stream.

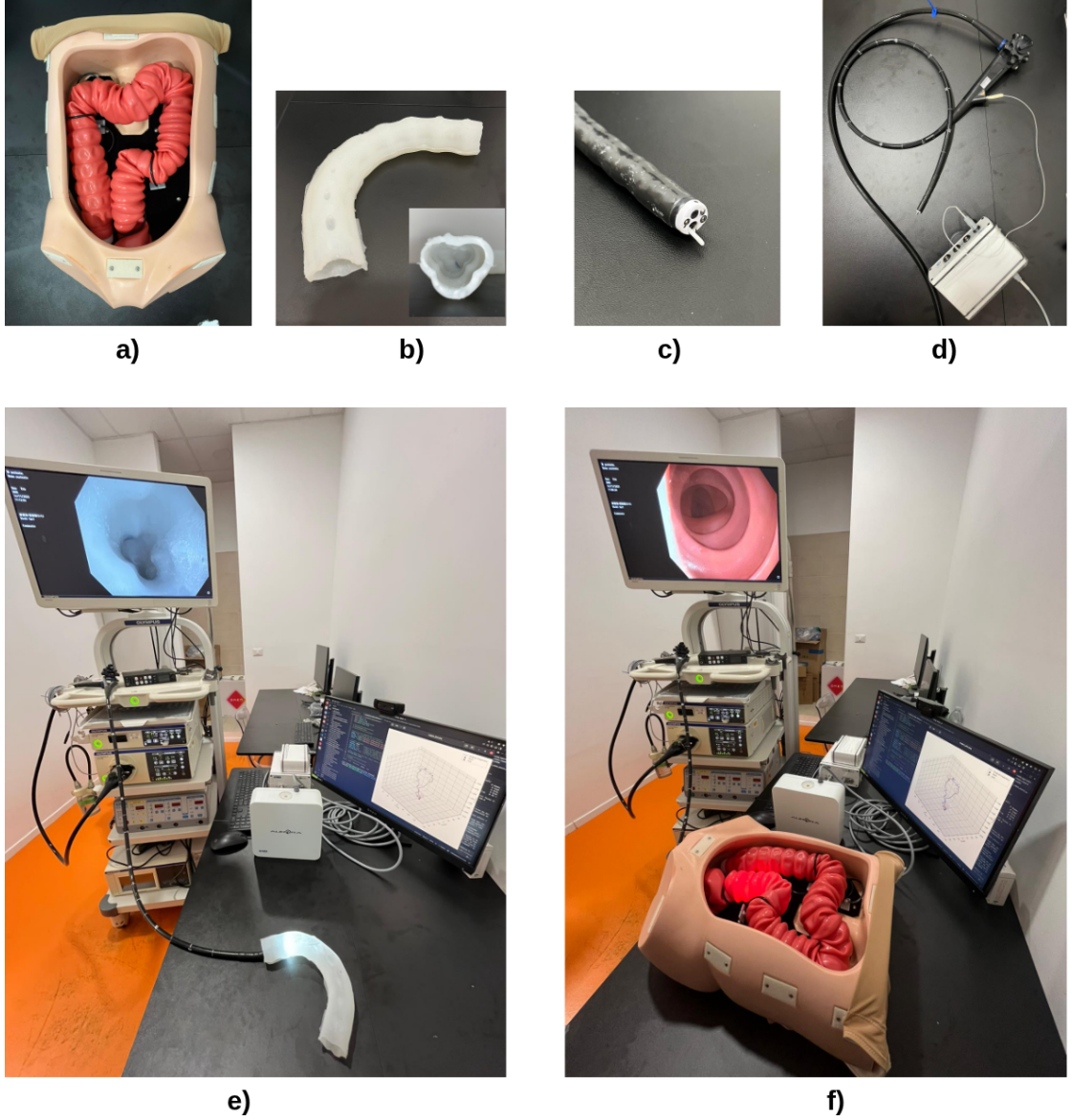


Figure 3.13: Aurora electromagnetic tracking experimental setup. (a) Kagaru phantom with realistic mucosal texture. (b) Silicone phantom segment. (c) 6-DoF electromagnetic sensor mounted inside the endoscope working channel. (d) Endoscope with sensor cable attached to the Sensor Interface Unit (SIU). (e) Complete acquisition setup with silicone phantom positioned within the Aurora tracking volume. (f) Acquisition setup with Kagaru phantom during data recording.

3.6.5 Endoscope camera calibration

The endoscopic camera was calibrated in order to obtain the intrinsic matrix required for 3D reconstruction. A checkerboard pattern of 7×8 squares (each square $1 \text{ cm} \times 1 \text{ cm}$) was used as the calibration target. The calibration process began by recording a video of the checkerboard from various angles and distances. From this video, around 20 frames were selected—each showing the pattern from a different orientation—to provide sufficient pose diversity for reliable calibration.

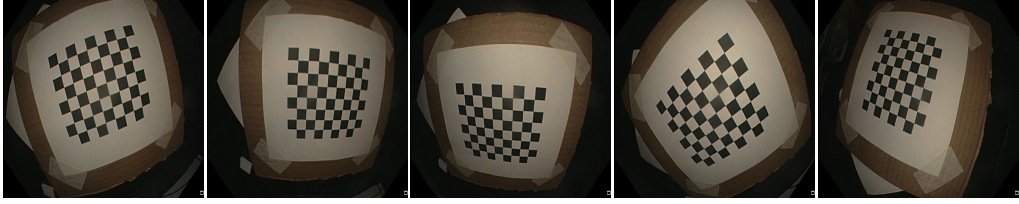


Figure 3.14: 5 of the 20 endoscope images used for calibration.

Following the procedure described in the OpenCV tutorial⁵, first the checkerboard corners were located in each selected image after converting the image to grayscale.

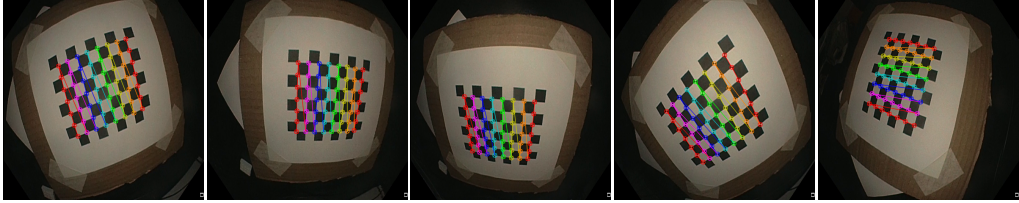


Figure 3.15: Detected checkerboard corners in the calibration images.

Then the 3D object points were defined on the checkerboard plane ($Z = 0$), with coordinates such as $(0,0,0)$, $(1 \text{ cm}, 0, 0)$, $(2 \text{ cm}, 0, 0)$, etc., consistent with the real-world square size. After collecting the corresponding image points and object points from all of the chosen frames, `cv2.calibrateCamera()` was used to compute the camera matrix (containing focal lengths f_x, f_y and principal point c_x, c_y), the distortion coefficients, and the rotation/translation vectors for each view. This intrinsic matrix was subsequently used in the 3D reconstruction pipeline to accurately map image points into the camera/world coordinate system.

⁵https://docs.opencv.org/4.x/dc/dbb/tutorial_py_calibration.html

Chapter 4

Results

4.1 TSDF validation

The TSDF-based reconstruction module was evaluated on the C3VD dataset, which provides shorter sequences with nearly static scenes and includes the ground-truth 3D models used to generate the phantom. This enabled a direct comparison between the reconstructed meshes and the corresponding GT surfaces for system validation. In addition to qualitative visual inspection as represented in Figure 4.1, quantitative metrics such as the surface overlap at $\tau = 4$ mm were computed for some sequences of the dataset, yielding a mean overlap of 96%.

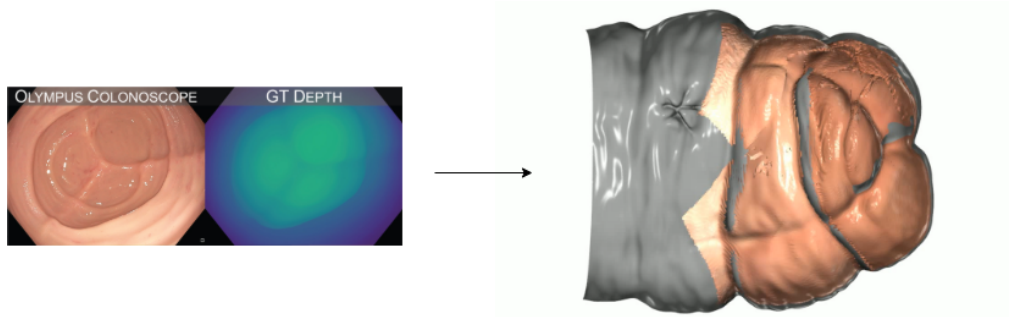


Figure 4.1: 3D reconstruction obtained from a C3VD video (colored) compared with the GT model (gray)

4.2 Depth prediction

To quantitatively evaluate the accuracy of the predicted depth maps, a dedicated evaluation pipeline was implemented. The script performs a systematic, frame-by-frame comparison between the estimated depth and the ground-truth depth provided by the *SimCol3D* dataset. For each frame, the depth map is loaded and converted into millimetres following the dataset specifications, where 16-bit depth images are normalised such that the maximum value (65535) corresponds to a physical distance of **200 mm**. This ensures consistency across models and allows all error metrics to be expressed in true physical units. When necessary, the predicted map is resized to match the ground-truth resolution, and a validity mask removes zero, undefined, or non-finite pixels to guarantee a robust numerical comparison. An example of input RGB frames together with the corresponding ground-truth depth and the depths predicted by different models is shown in Figure 4.2, highlighting the variability in quality across methods.

The evaluation focuses on a set of widely adopted depth-estimation metrics, each capturing a complementary aspect of prediction quality. The **Mean Absolute Error (MAE)** quantifies the average point-wise deviation in millimetres, providing an intuitive measure of reconstruction accuracy. The **Root Mean Squared Error (RMSE)** penalises larger deviations more heavily, making it sensitive to local prediction failures such as noise, depth discontinuities, or artefacts in specular regions. The **Relative Absolute Error (RelAbsErr)** evaluates the magnitude of the error relative to the true depth, which is particularly relevant in anatomical environments characterised by significant depth variation such as the colonic lumen.

In addition to these absolute error metrics, the pipeline computes the δ_1 , δ_2 , and δ_3 **accuracy thresholds**, expressing the percentage of pixels whose predicted depth lies within progressively relaxed multiplicative bounds of the ground truth (1.25 , 1.25^2 , 1.25^3). These metrics are especially important in endoscopic scenarios because they quantify how often the estimator remains within a physically acceptable deviation, regardless of absolute scale. High δ -values indicate globally reliable predictions, whereas lower values highlight the frequency and severity of under- or over-estimation errors.

Finally, the **Pearson correlation coefficient** is computed to capture the global linear relationship between predicted and ground-truth depth distributions. Unlike MAE and RMSE, correlation is insensitive to uniform scale shifts, thereby reflecting whether the model preserves the overall geometric structure of the scene even when absolute depth predictions are biased.

All metrics computed across the entire set of **15 sequences** of the *Synthetic-Colon_I* dataset are summarised in Table 4.1, providing a comprehensive quantitative benchmark of each model’s performance. Absolute error metrics assess geometric fidelity in millimetres, relative and tolerance-based metrics measure

robustness to depth variation, and correlation evaluates large-scale structural coherence.

Most importantly, because depth constitutes the **core input for the volumetric reconstruction pipeline**, its accuracy directly affects the quality, stability, and anatomical consistency of the reconstructed 3D colon geometry. Reliable depth predictions are essential for producing consistent TSDF volumes, reducing reconstruction artefacts, and preserving the morphology of haustral folds and lumen curvature.

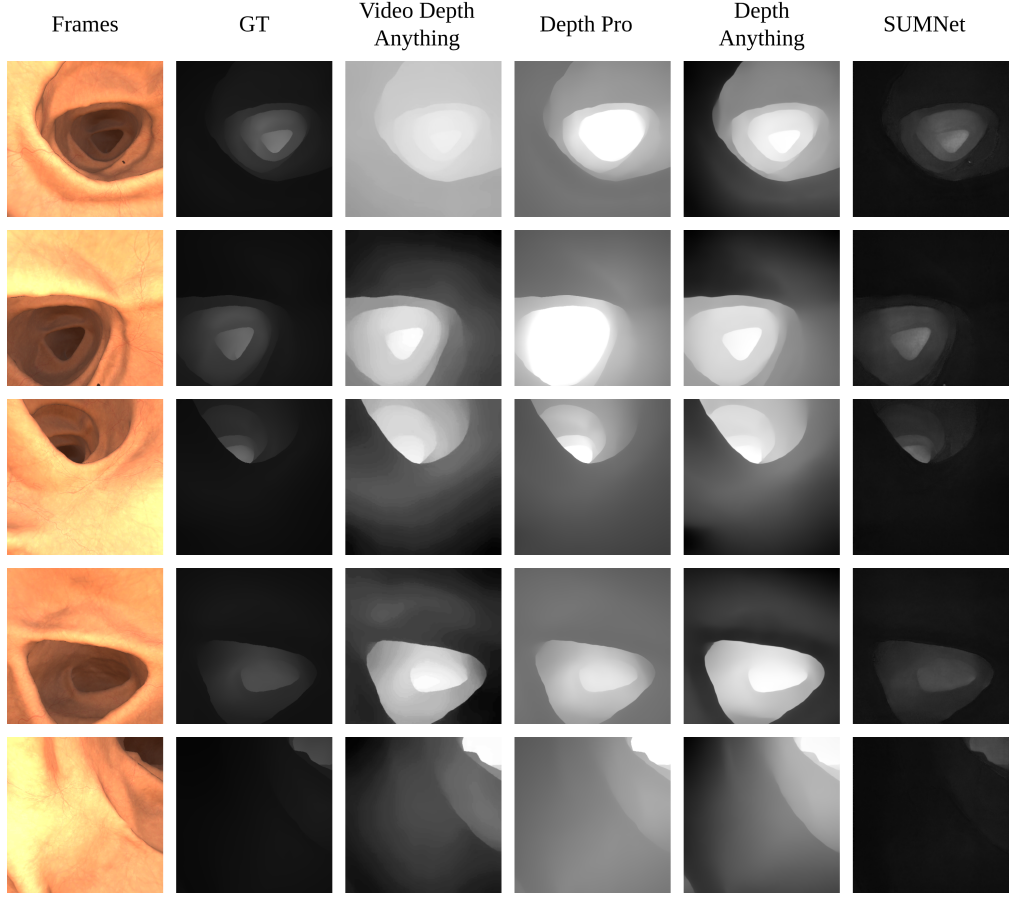


Figure 4.2: Examples of depth prediction on *SimCol3D* sequences.

Table 4.1: Depth estimation metrics on SimCol3D SyntheticColon_I (15 sequences). MAE and RMSE are both expressed in millimeters

Model	MAE	RMSE	RelAbsErr	δ_1 [%]	δ_2 [%]	δ_3 [%]	Corr.
Video-Depth-Anything	134.26	134.76	7.55	0.01	0.14	0.57	0.89
ML-Depth-Pro	121.78	124.36	6.39	0.01	0.10	0.42	0.81
Depth-Anything	78.93	87.51	3.66	1.21	2.80	5.34	0.89
SUMNet	2.04	2.94	0.11	90.91	98.33	99.48	0.98

4.3 Pose prediction

4.3.1 Quantitative pose estimation results

For each colonoscopy sequence of the SyntheticColon_I dataset, the bimodal pose network is evaluated. The system takes the RGB frames as input and, for each pair of frames at fixed stride, predicts the relative 6-DoF camera motion. These relative poses are then integrated into an absolute trajectory and compared against the ground-truth simulator trajectory as it is shown in Figure 4.3.

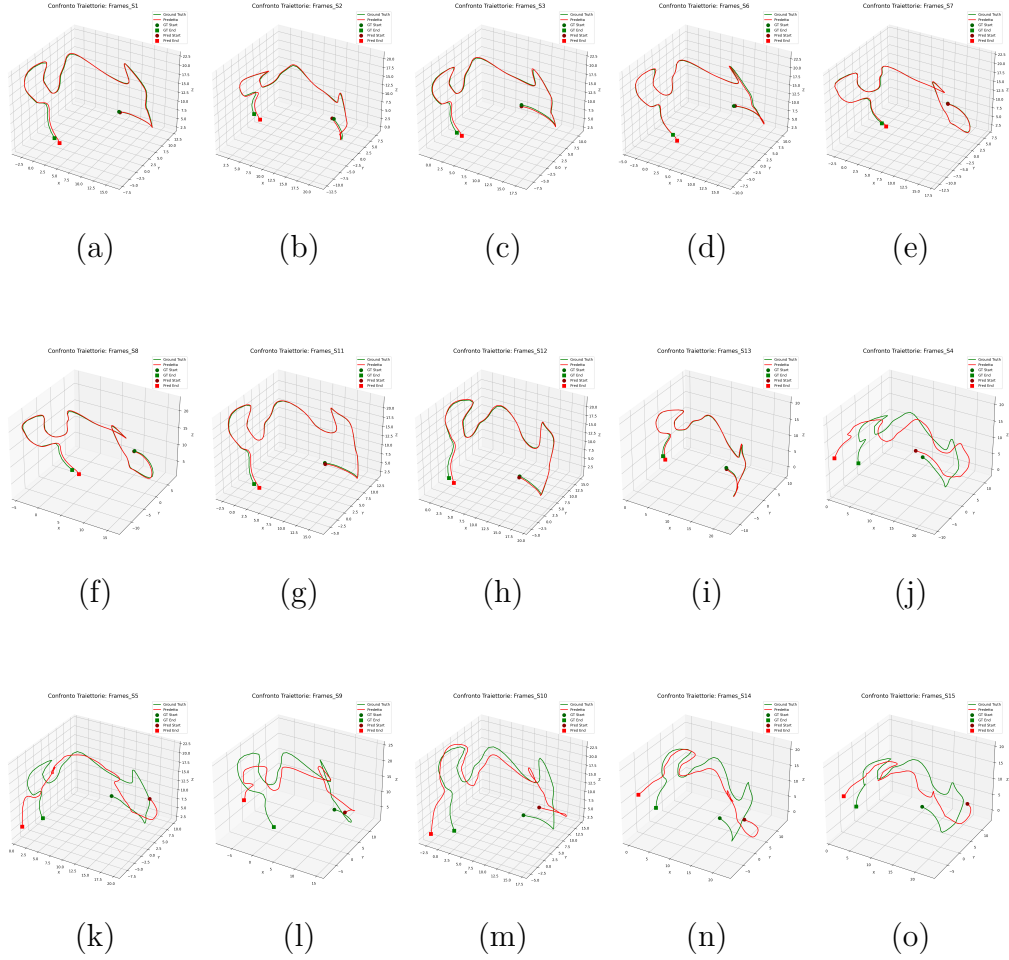


Figure 4.3: Predicted camera trajectories on SimCol3D sequences. The figure shows the spatial alignment between predicted (red) and ground-truth (green) trajectories across multiple sequences from SyntheticColon_I. Sequences (a)–(g) belong to the training split, (h)–(j) to the validation split, and (k)–(o) to the test split.

Let $\hat{\mathbf{T}}_j \in SE(3)$ and $\mathbf{T}_j \in SE(3)$ denote the predicted and ground-truth camera poses at frame j , and let $\hat{\mathbf{p}}_j, \mathbf{p}_j \in \mathbb{R}^3$ be their translation components. The script constructs both forward (insertion) and backward (withdrawal) trajectories by chaining the predicted relative motions; all quantitative metrics reported here are computed on the backward trajectory. A similarity-based alignment (rotation, translation and scale) between predicted and ground-truth trajectories is additionally estimated for 3D reconstruction purposes, but *all the pose metrics described below are computed on the raw predicted trajectory*, without applying any rigid alignment or rescaling. Before comparison, the predicted and ground-truth trajectories are simply trimmed to the same temporal length.

Absolute Trajectory Error (ATE). The Absolute Trajectory Error measures the positional discrepancy between predicted and ground-truth camera centres. For each frame index j ,

$$e_j^{\text{ATE}} = \|\hat{\mathbf{p}}_j - \mathbf{p}_j\|_2 \quad [\text{cm}].$$

From this sequence of errors, the script computes several summary statistics per sequence (mean, median, standard deviation, RMSE, minimum, maximum). In the subsequent analysis, we focus on the mean ATE,

$$\text{ATE}_{\text{mean}} = \frac{1}{N+1} \sum_{j=0}^N e_j^{\text{ATE}},$$

which is the quantity reported in the result tables for each data split.

Relative Pose Error (RPE). The Relative Pose Error characterises the local consistency of the odometry by comparing the relative motion between consecutive frames. For each j , the ground-truth and predicted relative transforms are

$$\mathbf{Q}_j = \mathbf{T}_j^{-1} \mathbf{T}_{j+1}, \quad \mathbf{P}_j = \hat{\mathbf{T}}_j^{-1} \hat{\mathbf{T}}_{j+1},$$

and the residual error transform is

$$\mathbf{E}_j = \mathbf{Q}_j^{-1} \mathbf{P}_j \in SE(3).$$

Writing

$$\mathbf{E}_j = \begin{bmatrix} \mathbf{R}_j^\Delta & \mathbf{t}_j^\Delta \\ \mathbf{0}^\top & 1 \end{bmatrix},$$

the translational and rotational RPE for step j are defined as

$$d_j^{\text{trans}} = \|\mathbf{t}_j^\Delta\|_2 \quad [\text{cm}], \quad d_j^{\text{rot}} = \arccos\left(\frac{\text{trace}(\mathbf{R}_j^\Delta) - 1}{2}\right) \quad [\text{rad}],$$

with d_j^{rot} converted to degrees as $d_j^{\text{rot,deg}} = d_j^{\text{rot}} \cdot 180/\pi$. Per sequence, the script computes mean, median and RMSE of both d_j^{trans} and $d_j^{\text{rot,deg}}$. In the tables we report the mean translational and rotational RPE,

$$\text{RPE}_{\text{mean}}^{\text{trans}} = \frac{1}{N} \sum_{j=0}^{N-1} d_j^{\text{trans}}, \quad \text{RPE}_{\text{mean}}^{\text{rot}} = \frac{1}{N} \sum_{j=0}^{N-1} d_j^{\text{rot,deg}}.$$

Drift of the positional error. To summarise how the positional error grows along the sequence, the system fits a simple linear model to the ATE values. Let e_j^{ATE} be as above and $j \in \{0, \dots, N\}$ the frame index. A least-squares line

$$e_j^{\text{ATE}} \approx aj + b$$

is fitted using `numpy.polyfit`, and the following quantities are extracted:

$$\text{drift_rate} = a \quad [\text{cm/frame}], \quad \text{drift_error_growth} = e_N^{\text{ATE}} - e_0^{\text{ATE}} \quad [\text{cm}].$$

The drift rate expresses the average increase in positional error per frame, while the error growth measures the total increase in ATE over the sequence. These two scalars are the drift-related metrics later aggregated across sequences.

Sequence-level metrics and split-wise aggregation. For each sequence in the dataset, the script computes the following metrics:

- mean Absolute Trajectory Error (ATE_{mean});
- mean translational Relative Pose Error ($\text{RPE}_{\text{mean}}^{\text{trans}}$);
- mean rotational Relative Pose Error ($\text{RPE}_{\text{mean}}^{\text{rot}}$);
- drift rate (`drift_rate`);
- total drift error growth (`drift_error_growth`).

These per-sequence metrics are then aggregated according to the data split. The aggregation is performed separately for the training sequences (S1, S2, S3, S6, S7, S8, S11), the validation sequences (S12, S13), and the test sequences (S4, S5, S9, S10, S14, S15). For each split, the mean value across all sequences is computed for every metric. The resulting split-wise statistics are reported in Table 4.2.

Table 4.2: Pose estimation results aggregated by data split.

Metric	Train	Validation	Test
ATE_{mean} [cm]	0.39	0.44	8.42
$\text{RPE}_{\text{mean}}^{\text{trans}}$ [cm]	0.01	0.01	0.09
$\text{RPE}_{\text{mean}}^{\text{rot}}$ [deg]	0.11	0.12	1.88
drift_rate [cm/frame]	0.001	0.002	0.07
drift_error_growth [cm]	0.82	0.97	21.77

4.4 Missing regions

One of the main goal of this thesis is to assess how much mucosa has been observed during a colonoscopic exam. To this end, after having reconstructed the 3D mesh of the scene, a Missing region analysis has been performed. The informations with the most clinical relevance are:

- **Percentage of unobserved surface:** the percentage of unobserved region.
- **Distribution of missing regions:** spatial mapping of unobserved areas.

The understanding of where blind spots occur can inform endoscopists about potential areas of concern, guiding more thorough examinations and improving overall diagnostic accuracy.

4.4.1 Percentage of unobserved surface

Using the Missing region analysis module presented in Chapter 3, 30 sequences from the SimCol3D dataset have been analyzed. The results are reported in Table 4.3 and Table 4.4.

Table 4.3: Unobserved regions for SyntheticColonI.

Sequence	Unobserved region [%]	Unobserved area [mm ²]
S1	20.9	37191
S2	21.7	37953
S3	25.0	42807
S4	21.1	34373
S5	21.3	35762
S6	18.9	34306
S7	18.6	33905
S8	17.9	33908
S9	20.0	36627
S10	18.0	32523
S11	19.1	34952
S12	19.1	34445
S13	17.6	30687
S14	19.2	34270
S15	18.9	32717
Mean	19.6±1.9	35095±2823

Table 4.4: Unobserved regions for SyntheticColonII.

Sequence	Unobserved region [%]	Unobserved area [mm ²]
B1	21.2	61199
B2	22.7	58936
B3	24.3	58878
B4	25.1	66104
B5	23.6	61483
B6	22.9	63915
B7	22.5	60987
B8	25.3	64116
B9	23.4	60373
B10	24.1	59729
B11	24.7	61345
B12	23.4	63103
B13	23.5	63940
B14	24.3	63534
B15	23.6	62382
Mean	23.6±1.1	62001±2100

4.4.2 Distribution of missing regions

The distribution of missing regions has been analyzed to identify specific areas where mucosa was not observed during the colonoscopic exam. This analysis provides valuable insights into the spatial characteristics of unobserved regions. Each missing region, derived from the comparison between the reconstructed mesh and the Poisson-reconstructed (closed) mesh described in Chapter 3, is geometrically defined as a set of connected faces represented in the coordinate frame of the reconstructed mesh. Therefore, knowing the spatial location of each missing region is essential for subsequent analyses. To this end, the position of every region was determined by computing the centroid of its connected faces, thus providing a compact geometric descriptor of its location within the reference frame of the reconstructed colon.

For each missing centroid identified, the coordinates and the distance from the start of the nearest point belonging to the camera trajectory is identified, in order to help endoscopists understand to which depth the endoscope must be reinserted to cover for the missed region.

Those information have been extracted for all the 30 sequences from the Sim-Col3D dataset.

Since the missing regions identification system is able to detect unobserved areas

down to a minimum threshold of 20 triangles and 5.0 mm² (as specified in Chapter 3), the resulting output can include a large number of small, geometrically isolated regions that may have limited clinical significance. While such fine-grained detection demonstrates the sensitivity of the pipeline, it also introduces the challenge of prioritizing among numerous candidates when providing actionable feedback to endoscopists.

To address this issue and ensure that the most clinically relevant information is highlighted, the system ranks all detected missing regions by their surface area and retains only the **top 20 largest unobserved regions** for detailed reporting. This filtering strategy is motivated by the observation that larger missing regions are more likely to represent anatomically significant blind spots—such as those behind haustral folds or in poorly illuminated concave areas—where the risk of overlooking pathological features is elevated. Conversely, very small unobserved patches often arise from transient occlusions, surface noise, or minor reconstruction artifacts, and thus carry less weight in the overall assessment of examination completeness.

For each of the 20 largest missing regions, the system provides the following descriptive data:

- The **centroid coordinates** in the reconstructed mesh reference frame, offering a precise geometric localization of the unobserved area.
- The **surface area** of the region, expressed in mm², which quantifies its extent and clinical importance.
- The corresponding **percentage of the total trajectory length** at which the nearest camera position occurs, allowing endoscopists to quickly identify the approximate depth of insertion required to revisit the unobserved zone.

An example of this structured output, generated for a single sequence (S4 from SyntethicColon-I), is presented in Table 4.5. The table format facilitates rapid interpretation and supports decision-making during or after the colonoscopic procedure, by directing attention to the most substantial coverage gaps in a prioritized and quantitatively informed manner.

Rank	Area [mm ²]	Centroid (x,y,z)	Nearest traj. point (x,y,z)	Traj. %
1	5423	(−256, −99, −168)	(−238, −137, −234)	86.7
2	3661	(−98, −172, −23)	(−108, −136, −51)	30.3
3	3191	(−120, −170, −187)	(−121, −164, −188)	68.2
4	2705	(−242, −138, −209)	(−241, −139, −234)	87.0
5	2314	(−100, −132, −91)	(−96, −134, −54)	29.1
6	1894	(−140, −129, −161)	(−131, −166, −189)	67.2
7	1249	(−153, −152, −36)	(−148, −130, −42)	34.2
8	1133	(−184, −166, −199)	(−170, −177, −189)	63.3
9	1103	(−191, −183, −19)	(−189, −157, −57)	39.8
10	1028	(−158, −164, −205)	(−156, −170, −191)	64.8
11	960	(−211, −159, −226)	(−222, −130, −233)	85.0
12	880	(−260, −114, −218)	(−245, −141, −234)	87.5
13	877	(−218, −158, −20)	(−221, −144, −38)	9.9
14	811	(−295, −165, −35)	(−293, −135, −48)	2.9
15	652	(−95, −193, −5)	(−109, −136, −51)	30.4
16	617	(−163, −199, −16)	(−156, −172, −64)	43.3
17	529	(−220, −153, −192)	(−226, −132, −233)	85.5
18	373	(−141, −215, −6)	(−150, −175, −66)	44.0
19	306	(−92, −213, 1)	(−115, −176, −78)	47.5
20	282	(−148, −194, −10)	(−153, −174, −65)	43.7

Table 4.5: Statistics of the 20 largest unobserved surface regions.

4.4.3 Uncertainty map of the unobserved regions

As described in Chapter 3, the heatmap was generated by summing the meshes corresponding to the missing regions from each of the 15 reconstructions. The resulting accumulation map highlights how frequently each surface area was classified as unobserved across the dataset. Although the combined map still presents holes—areas that were never included among the missing regions in any sequence—it provides an intuitive visual overview of coverage consistency. Brighter zones, corresponding to surfaces repeatedly identified as unobserved, cluster mainly around the major

mucosal folds and in the concave areas behind the haustral ridges of the colon, where visibility is most often lost due to self-occlusion and limited endoscopic field of view, as shown in Figure 4.4.

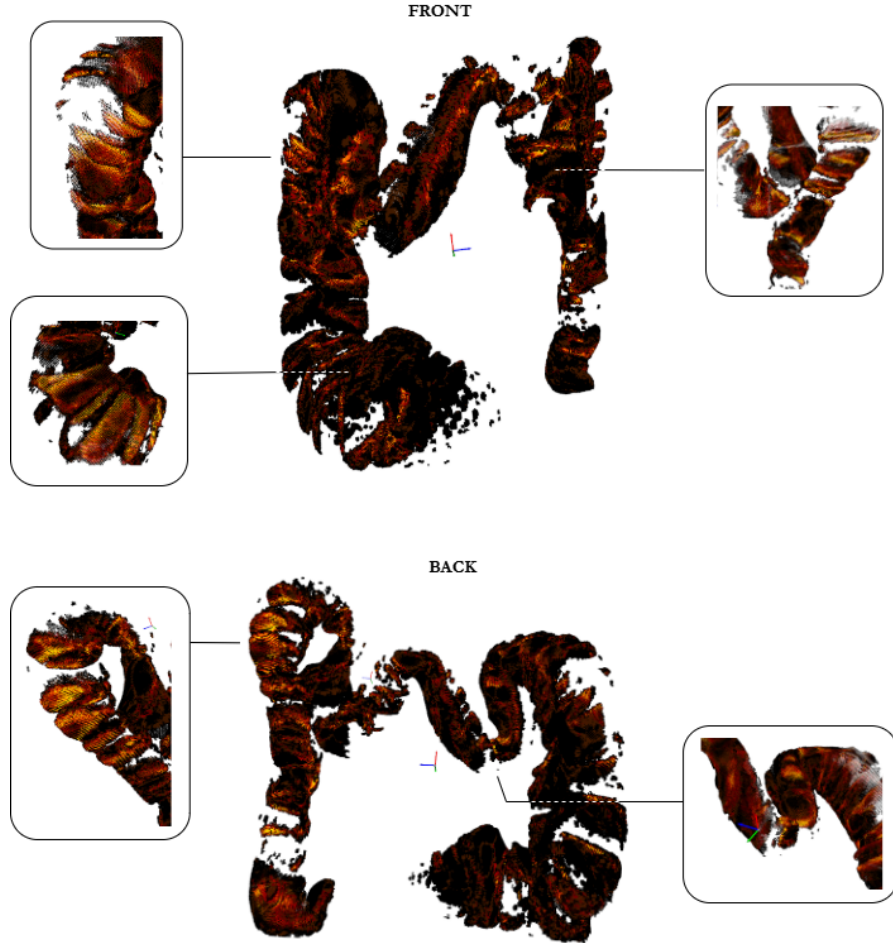


Figure 4.4: Heatmap of the missing regions across 15 sequences, the zoomed regions highlight the areas most frequently unobserved

4.5 Ablation study - Mesh to Mesh

The reconstruction pipeline has been evaluated on the *SyntheticColon_I* anatomy, specifically focusing on the reconstruction of surfaces for sequences S1–S15 under four input configurations:

- **GT Depth + GT Pose.**
- **Pred Depth + GT Pose.**
- **GT Depth + Pred Pose.**
- **Pred Depth + Pred Pose.**

For each configuration and sequence, a reconstructed test mesh $\mathcal{M}_{\text{test}}$ is compared against the ground-truth colon surface \mathcal{M}_{gt} provided by the dataset.

Alignment and distance computation. Before computing distances, the test mesh is rigidly aligned to the ground-truth mesh in $SE(3)$. The script first includes an optional pre-alignment step that can mirror the test mesh with respect to the plane containing the bisector between the Y and Z axes (normal $\propto (0, 1, -1)$). This is intended to compensate for possible axis flips between coordinate conventions; for the *SyntheticColon_I* reconstructions used in this study, this mirroring is disabled.

The core alignment is then performed via a robust coarse-to-fine registration pipeline operating on point clouds sampled from the two meshes:

- **Surface sampling for alignment:** Both meshes are first converted to point clouds by Poisson–disk sampling, with 50,000 points per mesh. These point sets are used exclusively for registration, while a denser sampling is employed later for metrics.
- **Outlier removal:** Statistical filtering is applied independently to both point clouds using a neighborhood size $k = 20$ and standard deviation ratio $\text{std_ratio} = 2.0$, removing isolated outliers that could destabilize RANSAC and ICP.
- **Normal estimation:** Normals are estimated on the cleaned point clouds using a hybrid KD-tree search with $\text{radius} = 2 \text{VOXEL_SIZE}$. In the script, $\text{VOXEL_SIZE} = 8.0$ mm is used for all registration scales.
- **Coarse initialization:** A rigid initial alignment is obtained via RANSAC-based feature matching using Fast Point Feature Histograms (FPFH). The feature radius is set to 5VOXEL_SIZE , the maximum correspondence distance to 1.5VOXEL_SIZE , $\text{ransac_n} = 3$, and the convergence criteria to $(10^5, 0.999)$.

- **Refinement:** Starting from the RANSAC transform, a point-to-plane ICP refinement minimizes the following objective:

$$\min_{R, \mathbf{t}} \sum_i \left(\mathbf{n}_i^\top (R\mathbf{x}_i + \mathbf{t} - \mathbf{y}_i) \right)^2,$$

where \mathbf{n}_i are normals on the ground-truth point cloud and $(\mathbf{x}_i, \mathbf{y}_i)$ are corresponding points. The resulting rigid transform is finally applied to $\mathcal{M}_{\text{test}}$.

In all experiments reported here, ICP alignment is enabled; distances are thus computed in the common aligned coordinate frame.

Surface sampling (mesh \rightarrow point cloud). To make distance computations independent of mesh tessellation density, each mesh is converted into a point cloud through Poisson-disk sampling. The script samples $N = 80,000$ points per mesh:

$$\mathcal{P}_{\text{test}} = \text{Poisson}(\mathcal{M}_{\text{test}}; N), \quad \mathcal{P}_{\text{gt}} = \text{Poisson}(\mathcal{M}_{\text{gt}}; N).$$

This sampling step is performed both before and after alignment, depending on the evaluation stage.

Bidirectional distances. Given the aligned point sets, the script computes nearest-neighbour distances in both directions using Open3D’s accelerated KD-tree routines:

$$d_{\text{test} \rightarrow \text{gt}}(\mathbf{p}) = \min_{\mathbf{q} \in \mathcal{P}_{\text{gt}}} \|\mathbf{p} - \mathbf{q}\|_2, \quad d_{\text{gt} \rightarrow \text{test}}(\mathbf{q}) = \min_{\mathbf{p} \in \mathcal{P}_{\text{test}}} \|\mathbf{q} - \mathbf{p}\|_2.$$

The resulting distance arrays are then summarized statistically.

Metrics reported. To quantitatively assess the geometric fidelity of each surface reconstruction, the following metrics are computed. Each metric captures a different aspect of reconstruction quality, and their combined use provides a comprehensive evaluation of both *local accuracy* and *global consistency*.

- **Mean distance.** Measures the average point-to-surface discrepancy. It is a global indicator of reconstruction accuracy and is sensitive to systematic bias (e.g. consistent surface shrinkage or expansion). A lower mean implies that, on average, the reconstructed surface lies close to the ground truth.
- **Median distance.** The median is more robust to outliers than the mean. It summarizes the “typical” reconstruction error and is particularly useful when isolated regions exhibit large misalignment (e.g. specular areas or depth discontinuities).

- **RMSE (Root Mean Squared Error).**

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_i d_i^2}.$$

By squaring the distances, RMSE penalizes large deviations more strongly, making it a suitable metric for detecting geometric drift and large-scale distortions—effects that are especially relevant when camera poses are predicted rather than ground truth.

- **95th percentiles.** These quantify the behaviour of the *worst* 5% of the surface. They highlight localized failure modes (e.g. folds, self-occlusions, topological ambiguities) while being more stable than the absolute maximum. Such statistic is essential in colon anatomy, where errors often cluster in specific regions such as sharp bends or collapsed lumen areas.
- **Mean Chamfer distance.**

$$\text{Chamfer}_{\text{mean}} = \mathbb{E}[d_{\text{test} \rightarrow \text{gt}}] + \mathbb{E}[d_{\text{gt} \rightarrow \text{test}}].$$

The Chamfer distance aggregates both directional errors into a single symmetric measure. It is widely used in 3D reconstruction benchmarks because it balances accuracy and completeness: a surface may fit the ground truth well in one direction while missing regions in the other. A low Chamfer distance indicates both precise alignment and consistent surface coverage.

- **Hausdorff distance.**

$$d_H = \max \left\{ \sup_{\mathbf{p} \in \mathcal{P}_{\text{test}}} \inf_{\mathbf{q} \in \mathcal{P}_{\text{gt}}} \|\mathbf{p} - \mathbf{q}\|_2, \sup_{\mathbf{q} \in \mathcal{P}_{\text{gt}}} \inf_{\mathbf{p} \in \mathcal{P}_{\text{test}}} \|\mathbf{p} - \mathbf{q}\|_2 \right\}.$$

The Hausdorff distance measures the *worst-case* discrepancy between the two surfaces. This metric is particularly relevant in medical reconstruction scenarios: even a small region with large geometric distortion may render the model unsuitable for navigation, simulation, or clinical measurement.

- **Coverage under threshold τ .** With $\tau = 4$ mm, this metric quantifies the percentage of points from the reconstructed surface that lie within a clinically acceptable tolerance from the ground truth.

$$\text{Cov}_\tau = \frac{1}{|\mathcal{P}_{\text{test}}|} \sum_{\mathbf{p}} \mathbf{1}[d(\mathbf{p}) \leq \tau].$$

Coverage complements the distance-based metrics by measuring how much of the reconstructed surface is *consistent* with the reference anatomy. This is critical for colon reconstruction, where global alignment may be good but large regions can still deviate due to pose drift or depth noise.

Taken together, these metrics provide a detailed view of the reconstruction quality: mean and median assess typical accuracy, RMSE and percentiles capture extremal deviations, Chamfer and Hausdorff evaluate shape consistency, and coverage quantifies the practical usability of the reconstruction.

Those metrics have been computed for the 6 test sequences of the Synthetic-Colon_I dataset (S4, S5, S9, S10, S14, S15) and they are reported in Table 4.6.

Table 4.6: Mesh-to-mesh comparison metrics for ablation study (averaged over 6 test sequences).

Metric	Predicted Depth	Predicted Pose	Predicted both
Mean distance [mm]	3.2 ± 0.7	23.0 ± 4.9	26.6 ± 11.1
Median distance [mm]	2.3 ± 0.3	7.5 ± 1.1	11.1 ± 2.6
RMSE [mm]	4.9 ± 2.1	35.4 ± 7.0	42.2 ± 16.5
95th percentile [mm]	8.3 ± 2.4	84.2 ± 17.5	97.2 ± 32.2
Chamfer distance [mm]	4.5 ± 0.9	38.2 ± 8.8	41.7 ± 18.1
Hausdorff distance [mm]	39.9 ± 22.7	126.3 ± 19.5	137.5 ± 21.4
Coverage @ 4 mm [%]	82.3 ± 3.6	34.7 ± 3.8	23.6 ± 4.4

Outputs and visual diagnostics. For each comparison, the script generates a color-coded mesh where vertex colors represent distance to the ground-truth surface using the Open3D “jet” colormap. To optimize visualization and avoid oversaturation by extreme outliers, the color scale is manually tuned based on the reconstruction scenario: for configurations involving **predicted poses** (GT Depth + Pred Pose, Pred Depth + Pred Pose), the maximum distance (red) is capped at **130 mm**, reflecting the larger spatial drift typically observed when camera localization is estimated rather than ground-truth. Conversely, for the **predicted depth** scenario (Pred Depth + GT Pose), where geometric errors are more localized and smaller in magnitude, the color range is capped at **15 mm**. These thresholds were determined experimentally to ensure that the color mapping effectively highlights meaningful reconstruction errors while preserving interpretability across different input conditions. Some representative examples are reported in Figure 4.5. An interactive visualization overlays the color-coded test mesh (error map) on top of the reference GT mesh, which is displayed in light gray, allowing for an immediate qualitative assessment of spatial deviations.

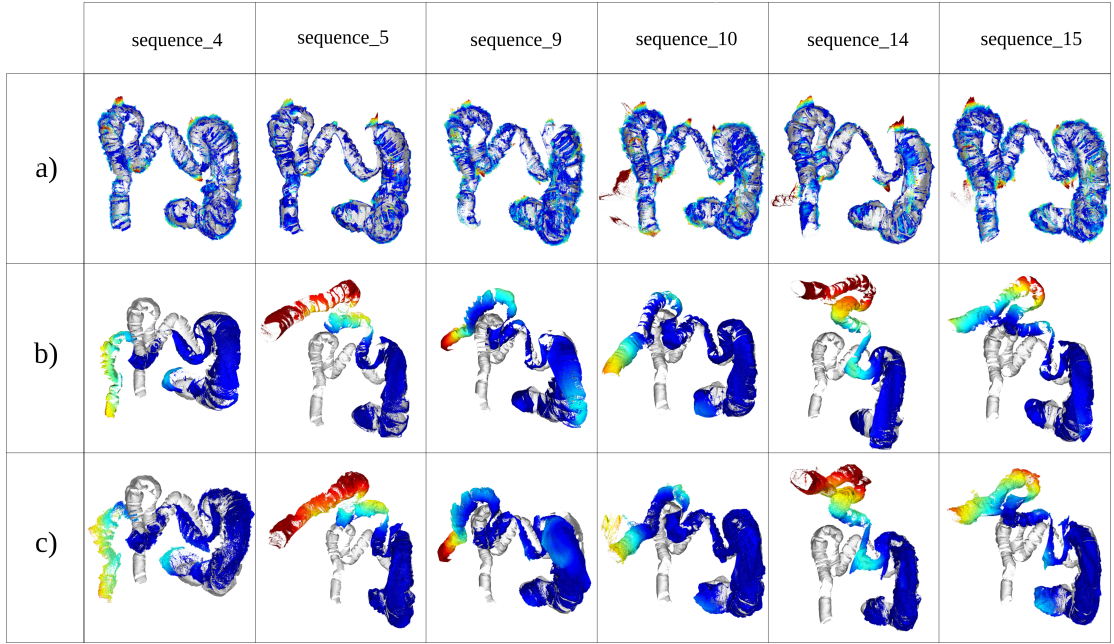


Figure 4.5: Mesh-to-mesh comparison for the ablation study. Each column corresponds to a different test sequence from the *SyntheticColon_I* dataset. Color encodes the point-to-surface distance from the reconstructed mesh to the ground-truth model (blue = low error, red = high error). (a) Ground-truth poses with **predicted depth maps**. (b) Ground-truth depth with **predicted poses**. (c) **Both depth and poses predicted**.

4.6 Aurora Dataset

Recorded data

The data obtained during the acquisition is reported in Table 4.7.

Table 4.7: Acquired dataset.

Sequence	Phantom	Frames	Pose	Duration [s]	Section
01	Kagaru	1779	1756/1779	59	Descending
02	Kagaru	2477	2435/2477	83	Sigmoid
03	Kagaru	786	781/786	27	Cecum
04	Silicone	709	702/709	24	/
05	Silicone	516	509/516	17	/
06	Silicone	491	489/491	16	/
07	Silicone	833	821/833	28	/
08	Silicone	735	721/735	24	/

The videos capture segments of the colon only, not full colonoscopies. As a result, spatial coverage is local and trajectories do not close the entire lumen; this should be considered when interpreting reconstruction accuracy and unseen-area metrics.

For some frames, the camera pose is missing due to AURORA EM-tracker signal dropouts or the probe temporarily leaving the tracking volume. To maintain 1:1 alignment between frames and poses, the pose files were interpolated so that the number of pose lines matches the number of frames (linear interpolation for translations; quaternion SLERP for rotations).

Sequences 1 to 3 were acquired on the Kagaru phantom, which has a more realistic mucosal texture and color but is more deformable. Sequences 4 to 8 were acquired on the Silicone partial phantom, which has a simpler texture but better simulates tissue stiffness and deformation qualities.

The silicone phantom sequences served mainly to test the acquisition setup under simpler, more controlled conditions, with the phantom’s transparency enabling external observation of the endoscope during acquisition.

The AURORA EM-tracker provides 6-DoF pose measurements at 40 Hz, while the endoscope camera captures video at 30 fps. To synchronize the two data streams, timestamps from both devices were aligned using a common reference signal, and the closest pose measurement in time was associated with each video frame. This ensures accurate temporal correspondence between the endoscopic images and their

respective camera poses for subsequent processing.

Predicted depths

The frames of each sequence were processed using **Depth-Anything**, the depth-prediction network introduced in Chapter 3. The resulting depth maps qualitatively show a coherent representation of the endoluminal geometry, with smooth transitions along the colon walls and consistent depth gradients across consecutive frames. However, a quantitative evaluation of the prediction accuracy was not possible, since the dataset was acquired experimentally and therefore does not include ground-truth depth maps for direct comparison. As a consequence, the assessment of the model’s performance in this context is limited to a qualitative inspection of the generated depth fields.

Calibration results

The calibration procedure described in Chapter 3 has been applied to the endoscope used for data acquisition, resulting in the intrinsic parameters reported in Table 4.8.

Table 4.8: Camera intrinsic parameters.

Parameter	Value
f_x	296.97 px
f_y	376.37 px
c_x	234.41 px
c_y	241.17 px

4.6.1 Dataset

In Figure 4.6, a sample of RGB frame and Depth map from each sequence of the dataset is shown, together with the whole trajectory .

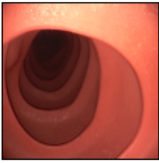
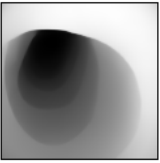
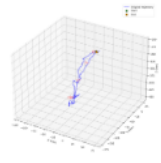
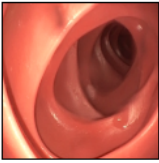
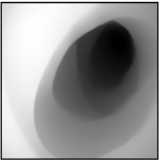
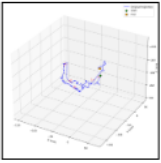
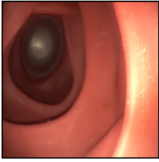
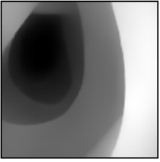
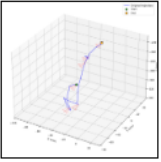


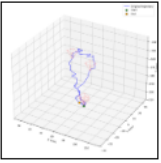
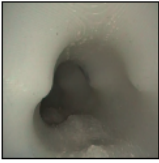

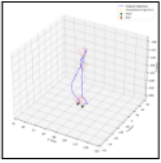
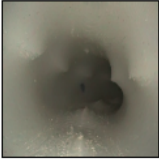
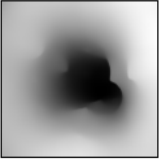
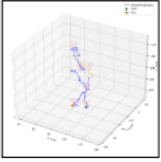
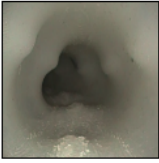
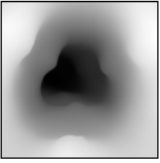
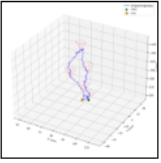
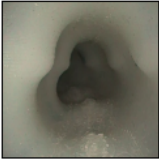

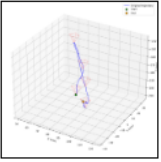
1			
2			
3			
4			
5			
6			
7			
8			

Figure 4.6: Sample frames from Aurora dataset.

Chapter 5

Discussion

5.1 Interpretation of Results

Depth Estimation

Accurate depth estimation is fundamental to the success of volumetric reconstruction, as it determines the geometric fidelity of the reconstructed colon surface. Among all tested models, the SUMNet model achieved the best quantitative performance in terms of L1, relative, and RMSE error, confirming its ability to adapt well to the SimCol3D colonoscopy domain. Its convolutional encoder–decoder architecture, combined with skip connections and max-unpooling, proved effective in preserving fine mucosal details while maintaining a regular and anatomically plausible tubular geometry.

By contrast, the foundation models considered in this work—Depth Anything, Video Depth Anything, and DepthPro—although highly successful for large-scale, metric depth estimation in natural scenes, showed clear limitations when transferred to the endoscopic setting. Their behaviour reflects the intrinsic characteristics of colonoscopy images, which frequently contain sparse or repetitive features, strong specular highlights, fluid reflections, low contrast regions, and non-Lambertian surfaces. These conditions differ substantially from those typically present in the training data of foundation models and introduce a marked domain gap. As a consequence, depth predictions become less accurate and less stable over time, and the resulting 3D reconstructions suffer from reduced geometric consistency.

The SUMNet model, on the other hand, delivered markedly better results. This improvement, however, is not surprising: the network was trained on data drawn from the very same SimCol3D dataset used for testing. Its performance therefore mainly reflects in-domain generalization rather than robustness to truly heterogeneous clinical conditions. In practice, SUMNet is the most effective depth predictor within the scope of the present experiments, but its behaviour on real

colonoscopy footage or on images with substantially different appearance remains uncertain and would require dedicated domain-adaptation strategies.

Pose Estimation

Pose estimation, performed using the bimodal regression network, produced globally consistent trajectories that were accurate enough to align consecutive frames and enable volumetric fusion. The network successfully captured the dominant forward-backward motion pattern typical of colonoscopy and generated pose sequences with sufficient stability for TSDF integration.

Quantitatively, the results show a clear distinction between the training/validation and test splits. On the training sequences (S1, S2, S3, S6, S7, S8, S11) and validation sequences (S12, S13), the network achieved excellent performance: mean ATE remained below 0.5 cm, translational RPE was around 0.01 cm per frame, and rotational RPE stayed below 0.12° . Visual inspection of the predicted trajectories (Figure 4.3) confirms this quantitative assessment, showing near-perfect overlap between predicted and ground-truth camera paths for these sequences. The estimated poses closely follow the true anatomical path of the endoscope throughout the entire sequence, with minimal drift accumulation even in longer acquisitions.

In contrast, performance on the test sequences (S4, S5, S9, S10, S14, S15) revealed more pronounced difficulties. Although the average metrics (mean ATE ≈ 8.4 cm, drift rate ≈ 0.07 cm/frame) remain within acceptable bounds for some applications, certain individual sequences exhibited substantial drift. In these cases, small per-frame pose errors compound progressively over time, causing the reconstructed trajectory to gradually deviate from the ground truth. This effect is particularly visible in Figure 4.3, where some test-set trajectories diverge significantly from the reference path as the sequence advances. Such drift compromises the geometric coherence of the estimated camera motion and, when propagated through volumetric fusion, can distort the global anatomical structure of the reconstructed colon.

These observations reflect the inherent challenge of monocular pose estimation in texture-poor or ambiguous environments, such as those encountered in longer sequences with limited parallax or in regions with repetitive mucosal patterns. The network’s strong in-domain performance on the training and validation sets demonstrates its ability to learn the colonoscopy motion model effectively, but the degradation on unseen test sequences highlights the difficulty of generalizing pose prediction to slightly different anatomical configurations or procedural variations, even within the same synthetic dataset.

As with depth, it is important to underline that the pose network was both trained and tested on SimCol3D sequences. The reported results therefore mainly quantify in-domain performance; the extent to which the same model would remain

stable on real colonoscopy videos, characterised by different textures, illumination, deformation patterns, and acquisition devices, is still unknown. It is reasonable to expect some degradation in predictive stability when moving away from the synthetic domain.

In addition, the methodological landscape for pose estimation is less mature than that for depth. While monocular depth prediction has benefited from large-scale pretraining and foundation models, learning-based camera pose estimation remains a narrower research area. Pose networks typically rely on relatively small and specialised datasets and often exhibit limited generalization outside their training distribution. This scarcity of broad, diverse training corpora, together with the intrinsic ill-posedness of monocular motion estimation in texture-poor environments, makes pose prediction particularly vulnerable to domain shift and helps explain the drift accumulation observed in more challenging or low-texture regions.

Missing Region Quantification

The quantitative analysis of missing regions showed that, across the 30 simulated colonoscopies in the *SyntheticColon_I* and *SyntheticColon_II* models, the percentage of unobserved mucosal surface averaged **19.6%** and **23.6%**, respectively.

This part of the study constitutes the central contribution of the thesis. The proposed missing-region identification pipeline is, to the best of our knowledge, the first to provide a detailed, geometry-based, mesh-level quantification of unseen mucosa directly from volumetric 3D reconstructions. The framework does not merely display unobserved regions qualitatively; it estimates spatially resolved coverage metrics that can be interpreted in clinical terms. This opens the door to a new form of objective, spatially explicit assessment of colonoscopy quality, allowing clinicians to understand, quantify, and potentially optimise the thoroughness of mucosal inspection.

The average percentages of unobserved mucosa are not only internally consistent across simulated sequences, but also align well with values reported in clinical literature, where approximately **22%–28%** of the colonic surface is estimated to remain unseen during routine colonoscopy, even when procedures are performed by experienced endoscopists under good bowel preparation. These figures, obtained from tandem colonoscopy meta-analyses and coverage studies, lend physiological plausibility and clinical relevance to the results obtained on SimCol3D.

The spatial distribution of missing regions provides additional insight. Inspection of both the individual reconstructed meshes and the aggregated heatmap across all sequences reveals a clear and clinically intuitive pattern: unseen areas cluster mainly behind the natural folds of the colon (haustra, or haustral ridges) and in the deep concave pockets produced by their overhang. Further blind spots appear along irregular mucosal relief and in small recesses shielded by slight protrusions or

villous-like surface undulations. These are exactly the anatomical configurations known to reduce visibility in forward-looking monocular endoscopy and are widely recognised in clinical practice as typical locations for missed lesions. The agreement between the reconstruction-based analysis and clinical experience supports the validity of the proposed coverage metrics.

Ablation Study Findings

The ablation study was designed to separate the individual contributions of depth and pose estimation to the overall reconstruction quality. By progressively replacing ground-truth inputs with predicted ones, it becomes possible to assess how each component influences both local geometric accuracy and global anatomical consistency.

When depth was predicted but ground-truth poses were used, the resulting reconstructions remained close to the reference surfaces. This behaviour reflects the good performance of the SUMNet model, whose depth predictions had already shown low error in direct comparisons with ground-truth depth. Within the SimCol3D domain, the network produced geometrically coherent surfaces, with only modest deviations from the reference. Local artefacts were occasionally visible along the mucosa—typically linked to depth errors in difficult regions—but these remained spatially confined and did not disrupt the global structure of the colon.

In contrast, replacing ground-truth poses with predicted ones, while still using ground-truth depth, led to a much more pronounced degradation in reconstruction quality. Although the depth maps accurately described local geometry, the errors in the estimated camera trajectory accumulated over time. Because alignment is anchored at the beginning of each sequence (near the cecum), even small pose inaccuracies at each frame propagate and compound along the trajectory, gradually displacing the reconstructed colon from the ground-truth anatomy. This is consistent with the general difficulty of monocular pose estimation, especially in the absence of loop closure and in sequences with limited parallax. In colonoscopy, such drift can severely affect the anatomical reliability of the reconstruction, causing the virtual model to no longer match the actual spatial configuration of the observed colon.

The fully estimated configuration, in which both depth and pose were predicted, combined the weaknesses of both components: local geometric noise induced by depth inaccuracies and global misalignment due to pose drift. The resulting meshes still captured the overall tubular morphology of the colon, but exhibited increased roughness, shape distortion, and larger point-to-surface discrepancies. These outcomes illustrate how challenging it is to perform accurate volumetric fusion when both geometric and motion cues are affected by predictive errors.

Taken together, the ablation results indicate that the TSDF-based pipeline

tolerates moderate depth inaccuracies reasonably well, but is considerably more sensitive to pose errors. Camera motion estimation emerges as the critical factor for reconstruction stability and anatomical fidelity, while depth quality primarily affects local surface detail rather than overall structural correctness.

More broadly, the study suggests a key design principle: reliable depth prediction is necessary to capture fine-scale mucosal morphology, but the success of the entire reconstruction process ultimately depends on the quality of pose estimation. Depth errors tend to remain local and generate limited artefacts, whereas pose inaccuracies accumulate over time and can irreversibly distort the global anatomy. To move toward clinically reliable, anatomically faithful 3D reconstructions, future work should therefore prioritise more robust pose estimation—through improved architectures, explicit drift-correction mechanisms such as loop closure, or hybrid learning–geometric formulations. Strengthening this component is essential to unlock the full potential of the pipeline and to approach dependable deployment in real colonoscopic practice.

Phantom Dataset Validation

The acquisition of the phantom dataset using the NDI Aurora electromagnetic tracking system was an important step toward bridging synthetic experiments and real-world application. The experimental setup was specifically designed to test whether the reconstruction framework could operate under realistic acquisition conditions. By synchronising RGB frames with externally measured camera poses, the system removes the need for learning-based pose estimation and leaves depth prediction as the only component to be inferred from images. Methodologically, this configuration allows the pipeline to be evaluated in a near-clinical setting, provided that a suitable, high-performance *metric* depth estimation model for endoscopy is available.

At present, however, such a depth model is still lacking. The depth maps obtained with Depth Anything are inherently *relative* and their precision is not sufficient for volumetric fusion. Consequently, the attempts to reconstruct the phantom surface did not lead to geometrically reliable meshes. This outcome underlines a central limitation of current depth estimation methods for real endoscopy: relative depth may suffice for coarse geometric reasoning, but physically meaningful 3D reconstruction requires metric depth with high spatial accuracy.

Despite these difficulties on the depth side, the Aurora system itself performed very well as a pose acquisition device. It delivered stable, real-time 6-DoF trajectories with minimal dropout and good temporal consistency, demonstrating its suitability for synchronous RGB–pose acquisition. These characteristics make electromagnetic tracking a valuable tool for generating high-quality ground-truth trajectories and for building hybrid acquisition pipelines in which only depth is

predicted from images.

Overall, the phantom experiment marks a significant milestone in shifting the reconstruction framework from pre-existing simulated datasets toward intraoperative acquisition. Although further advances in metric depth estimation are needed before full 3D reconstruction on real endoscopic footage becomes feasible, the successful integration and validation of the Aurora tracking system provide a solid basis for future work on clinically deployable 3D colon reconstruction.

5.2 Clinical Implications and Potential Applications

Although the experiments in this thesis are based primarily on synthetic and phantom data, the underlying motivation is explicitly clinical. Current colonoscopy practice still relies almost entirely on the endoscopist’s subjective perception of coverage and on indirect quality indicators such as withdrawal time or bowel preparation scores. In this context, a reconstruction framework capable of quantifying which regions of the mucosa have actually been seen offers a fundamentally different, spatially explicit view of examination quality.

The proposed coverage and missing-region metrics could, in principle, be used in several ways. First, they provide an objective measure of how much of the colonic surface is inspected under given procedural conditions, endoscopic techniques, or training levels, thereby complementing traditional quality indicators such as the Adenoma Detection Rate. Second, by localising blind spots along the colon, the method can highlight anatomically challenging regions, for instance haustral folds or sharply curved segments, where targeted training or modified withdrawal strategies may be most beneficial. Third, in a future real-time implementation, the same metrics could support intra-procedural feedback, warning the operator about under-inspected segments and potentially reducing the risk of missed lesions.

It should be stressed that these applications remain prospective and would require extensive validation on real patient data, as well as careful integration into clinical workflows. Nevertheless, the results obtained in simulation and on phantom acquisitions suggest that 3D reconstruction and coverage analysis have the potential to move colonoscopy assessment from global, indirect proxies toward a more direct and spatially grounded notion of quality.

5.3 Methodological Reflections and Design Choices

The design of the proposed pipeline reflects a series of methodological trade-offs. The choice of a TSDF-based volumetric representation, for example, favours robustness and watertight surface reconstruction at the expense of memory consumption

and fine-scale detail. Alternative approaches based on explicit point clouds or neural implicit fields could in principle offer higher resolution or more compact representations, but often at the cost of increased complexity, reduced interpretability, or more demanding training procedures. In this thesis, the emphasis was placed on a conceptually transparent and well-understood reconstruction backbone, which facilitated the development and analysis of the missing-region quantification module.

A similar compromise underlies the decision to work with monocular depth and pose estimation rather than relying on multi-view or stereo hardware. Monocular endoscopy remains by far the most common clinical scenario, and methods that operate under this constraint are more easily transferable to existing practice. At the same time, monocular cues are intrinsically weaker and more ambiguous than binocular or structured-light information, which partly explains the sensitivity of the pipeline to pose errors and the difficulty of obtaining truly metric depth. The use of synthetic data from SimCol3D further reflects a balance between realism and controllability: although the domain gap with real colonoscopy is non-negligible, simulation makes it possible to access ground-truth geometry and pose, systematically vary acquisition conditions, and isolate specific failure modes of the pipeline.

These design choices do not represent definitive answers but rather a pragmatic starting point. They make the problem tractable and allow for a clean experimental analysis, while at the same time highlighting where more sophisticated models or alternative sensing strategies may be most beneficial in future work.

5.4 Limitations

Several limitations emerged over the course of this study, reflecting both methodological constraints and the intrinsic challenges of endoscopic 3D reconstruction.

Computational and System-Level Limitations

A first limitation concerns computational performance. The current framework does not yet operate in real time: depth estimation, pose regression, TSDF fusion, mesh extraction, and missing-region computation together require processing times that are incompatible with intraoperative use. At this stage, the system is therefore more suited to post-acquisition analysis than to real-time guidance or quality assessment during the procedure. Achieving real-time operation would require substantial optimisation across all components, including model inference, volumetric integration, and mesh processing.

Limitations of the Synthetic Training Domain

Most experiments were carried out on the SimCol3D dataset which—although anatomically plausible—remains a synthetic environment with several important differences from clinical reality. The colon model in SimCol3D is static and quasi-rigid, whereas real colorectal tissue is highly deformable and continuously affected by insufflation, peristalsis, instrument manipulation, and patient motion. Real colonoscopy scenes also contain elements absent from the simulation, such as mucus, residual stool, fluid films, foam, smoke, and strong specular glare, all of which increase visual complexity and can destabilise depth or pose estimation. The trajectories in SimCol3D follow relatively smooth, predominantly forward paths, whereas real procedures often include retroflexion, torsion, abrupt reorientations, and complex looping. These discrepancies introduce a substantial domain gap that limits the ecological validity and generalisability of the results.

Depth and Pose Prediction Limitations

The depth and pose predictors form the main bottlenecks for reconstruction quality. The foundation models evaluated here performed poorly out-of-the-box on endoscopic imagery, confirming that they cannot be directly applied to colonoscopy without careful domain adaptation. The SUMNet model achieved strong performance, but this success is closely tied to training on SimCol3D and does not guarantee similar behaviour on real colonoscopy data. The same observation holds for the pose estimator, whose accuracy decreased notably in regions with sparse or ambiguous visual cues. In real data, depth prediction remained non-metric, and pose estimation suffered from cumulative drift in the absence of loop closure or global trajectory optimisation. These factors are particularly detrimental for long sequences and undermine the reliability of reconstructions outside the synthetic domain.

Limitations of Electromagnetic Tracking

The phantom experiments also exposed practical constraints of the NDI Aurora electromagnetic tracking system. Although Aurora is capable of delivering accurate, real-time pose measurements with few dropouts, it requires inserting a dedicated sensor into the endoscope’s instrument channel, which may not be compatible with all clinical workflows or instruments. Furthermore, electromagnetic tracking only operates correctly within the calibrated field of the emitter; if the scope moves outside this volume, accuracy degrades or the signal is lost altogether. These limitations restrict the feasibility of Aurora as a general-purpose solution for pose acquisition in routine colonoscopy.

5.5 Future Work

Future work will aim to improve both the accuracy and the practical usefulness of the proposed framework, with the long-term goal of achieving reliable, real-time 3D reconstruction during colonoscopy. On the basis of the results obtained in this thesis, several directions appear particularly relevant:

- **Improved depth and pose estimation toward metric, clinically usable predictions.** A first line of development concerns the systematic testing of alternative depth and pose estimation architectures, including recent foundation models, transformer-based networks, and hybrid geometric-learning methods. The objective is to obtain stable, drift-resistant, and fully *metric* predictions. In particular, depth estimators specifically designed or adapted for endoscopic imagery are required, so that the resulting depth maps no longer need ad-hoc rescaling and can be used directly for volumetric fusion on real data.
- **Progressive transition from synthetic to real datasets.** Although Sim-Col3D has been an effective testbed for algorithm design, future experiments should increasingly rely on more realistic data. This includes additional synthetic datasets with higher visual complexity, phantom acquisitions under different conditions, and, ultimately, real colonoscopy videos. Such a gradual shift is necessary to understand how each pipeline component behaves under realistic noise, tissue deformation, fluids, and illumination variability.
- **Hybrid SLAM-based systems for real-time reconstruction.** To move from offline reconstruction to intraoperative use, the current batch-oriented implementation will need to be complemented or replaced by hybrid SLAM frameworks. These systems would combine deep priors for depth and pose with classical geometric tracking, loop closure, and drift correction. Incremental TSDF fusion or alternative real-time mapping back-ends could then provide continuously updated reconstructions and coverage estimates throughout the procedure.
- **Integration with real-time polyp detection and segmentation.** Another natural extension is the integration of the reconstruction pipeline with deep-learning models for polyp detection and semantic segmentation. Projecting detected lesions onto the reconstructed 3D surface would enable spatially consistent reporting, support lesion follow-up, and open the way to applications such as risk maps of polyp distribution or navigation cues toward suspicious or under-inspected regions.
- **Richer clinical descriptors derived from missing regions.** The missing-region analysis can be further exploited beyond the global percentage of

unobserved mucosa. Additional descriptors could include the size distribution of blind spots, their distribution along specific colon segments, or their spatial relationship with detected lesions. These quantities could then be correlated with clinical indicators (e.g. Adenoma Detection Rate, withdrawal time, bowel preparation quality) to assess their potential as objective markers of examination quality.

- **Non-rigid and deformation-aware reconstruction models.** Future studies should also explore colon-specific non-rigid reconstruction approaches, such as deformation-aware TSDF formulations, scene-flow-based warping, or neural implicit representations capable of modelling tissue motion. Accounting for insufflation, peristalsis, and scope-induced deformation would substantially increase the anatomical realism of the reconstructed models.
- **Enhanced multimodal datasets and calibration.** The hybrid RGB-pose dataset acquired with the Aurora system could be enriched with structured-light, stereo endoscopy, or depth-probe acquisitions to provide ground-truth 3D information for quantitative validation. More sophisticated calibration procedures between the tracking system and the camera, together with robustness studies under different acquisition setups, would further improve the reliability of such datasets.
- **Interactive tools for real-time coverage monitoring.** Finally, the coverage and missing-region metrics proposed in this work could be incorporated into interactive visualization tools that provide endoscopists with intuitive feedback on which mucosal regions have been adequately inspected and which remain insufficiently explored. If combined with low-latency tracking and reconstruction, these tools could evolve into real-time decision-support systems during colonoscopy.

Overall, the discussion confirms that the proposed system can transform standard 2D endoscopic video into a coherent 3D representation that quantifies mucosal coverage. Although substantial work remains before clinical deployment is possible, the results obtained here provide a solid methodological basis for spatially aware colonoscopy and for the development of future real-time tools for objective quality assessment.

Chapter 6

Conclusions

6.1 Summary

The work presented in this thesis started from a practical question: whether it is possible to turn standard monocular colonoscopy video into a three-dimensional representation of the colon, and to use this representation to quantify which regions of the mucosa are actually observed during an examination.

The first step toward this goal was to identify and implement a 3D reconstruction strategy that could operate on existing datasets such as SimCol3D and C3VD, where ground-truth depth and pose are available. To this end, a volumetric pipeline based on a Truncated Signed Distance Function (TSDF) was developed. Starting from RGB frames and associated ground-truth camera trajectories, depth maps were integrated over time into a global TSDF volume, from which smooth, watertight meshes of the colonic lumen were extracted. This established a reference reconstruction framework against which the effect of different input sources could be systematically evaluated.

In a second phase, the focus shifted to the more challenging and realistic scenario in which only the video stream is available. To approximate this setting, several depth estimation methods and a bimodal camera pose estimation network were explored, trained, and evaluated on SimCol3D and related data. The study considered both large pre-trained depth models and a tailored, SUMNet architecture trained directly on colonoscopy-like images, alongside a deep network adapted to predict 6-DoF camera trajectories from image sequences. By analysing multiple error metrics and visualising the resulting reconstructions, the thesis assessed how well each approach could replace ground-truth inputs within the TSDF pipeline, both in isolation and in combination. An ablation study systematically combined ground-truth and estimated depth and pose in different configurations to disentangle their influence on reconstruction quality, clarifying to what extent inaccuracies in

each module degrade local geometric detail, global anatomical consistency, and the distances between reconstructed and reference surfaces.

In addition, a missing-region identification and coverage quantification method was introduced by closing TSDF meshes via Poisson surface reconstruction and comparing the original (open) and completed (closed) surfaces to derive spatially explicit coverage metrics. Applied to 30 simulated colonoscopies, this analysis showed that approximately 19–22% of the mucosal surface remains unobserved, aligning with clinical reports and indicating that such reconstruction-derived coverage measures can complement traditional quality indicators.

Finally, to move beyond purely synthetic data and test the feasibility of the approach in more realistic conditions, a dedicated experimental acquisition setup was designed. Using an Olympus colonoscope and an NDI AURORA electromagnetic tracking system, RGB videos on colon phantoms were recorded together with synchronised 6-DoF tip poses. This experimental configuration effectively removes the need for learned pose estimation and leaves depth as the only missing modality to be inferred, bringing the overall framework one step closer to the target scenario of reconstruction from video alone in a real endoscopic environment.

Across these stages, the thesis gradually progressed from controlled TSDF reconstructions based on ideal ground-truth inputs, to reconstructions driven by learned depth and pose predictors, and finally to an experimental acquisition setting that emulates key aspects of clinical use.

6.2 Concluding Remarks

This thesis has explored how monocular colonoscopy video can be turned into a three-dimensional, spatially aware representation of the colon, and how such a representation can be used to quantify which regions of the mucosa are inspected during an examination. The results obtained on synthetic and phantom data indicate that this goal is technically achievable and potentially relevant from a clinical perspective.

Methodologically, the work highlights several key points. First, while accurate depth estimation is important for preserving local surface detail, the stability and anatomical fidelity of the reconstruction depend primarily on camera pose. Depth errors largely remain local and produce limited artefacts, whereas pose inaccuracies accumulate over long sequences and can significantly distort the global geometry. Any attempt to move toward clinical deployment will therefore have to place strong emphasis on robust, drift-resistant pose estimation, likely combining learning-based priors with geometric optimisation and loop closure.

Second, the missing-region analysis introduced in this thesis shows that coverage can be measured and visualised in a way that aligns with anatomical intuition and

with patterns reported in the literature. The agreement between simulated findings and clinical estimates of unseen mucosa suggests that reconstruction-based coverage metrics could complement classical quality indicators such as adenoma detection rate, cecal intubation rate, and withdrawal time, providing a more spatially explicit view of how thoroughly the colon has been inspected.

Third, the phantom experiments demonstrate both the feasibility and current limitations of moving beyond synthetic data. The successful integration of electromagnetic tracking with clinical endoscopes shows that accurate RGB-pose acquisition is achievable in realistic conditions, reducing the problem to depth estimation alone. However, the limited performance of existing depth models on real endoscopic imagery confirms that robust depth estimation in flexible endoscopy remains an open challenge.

Looking ahead, key directions include extending the pipeline to real patient data, incorporating non-rigid reconstruction for tissue deformation, and integrating real-time polyp detection. If these elements can be combined, 3D reconstruction and coverage analysis may evolve from an offline research tool into a practical component of everyday colonoscopy.

In conclusion, while significant challenges remain before clinical deployment, this thesis demonstrates that geometry-aware, quantitative assessment of colonoscopy quality is within reach. By making visible what has and has not been seen, the proposed framework points toward more objective and spatially informed colorectal cancer prevention.

Bibliography

- [1] Usman Khan, AmanUllah Yasin, Muhammad Abid, Imran Shafi, and Shoab A. Khan. «A Methodological Review of 3D Reconstruction Techniques in Tomographic Imaging». In: (2018) (cit. on p. 9).
- [2] Johannes L. Schonberger and Jan-Michael Frahm. «Structure-From-Motion Revisited». In: (June 2016) (cit. on p. 12).
- [3] Aji Resindra Widya, Yusuke Monno, Masatoshi Okutomi, Sho Suzuki, Takuji Gotoda, and Kenji Miki. «Whole Stomach 3D Reconstruction and Frame Localization From Monocular Endoscope Video». In: *IEEE Journal of Translational Engineering in Health and Medicine* (2019) (cit. on p. 12).
- [4] Pål Anders Floor, Ivar Farup, and Marius Pedersen. «3D Reconstruction of the Human Colon From Capsule Endoscope Video». In: *IEEE Access* (2025) (cit. on p. 12).
- [5] H. Durrant-Whyte and T. Bailey. «Simultaneous localization and mapping: part I». In: *IEEE Robotics & Automation Magazine* (2006) (cit. on p. 13).
- [6] Raúl Mur-Artal and Juan D. Tardós. «ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras». In: *IEEE Transactions on Robotics* (2017) (cit. on p. 14).
- [7] Kailing Wang, Chen Yang, Yuehao Wang, Sikuang Li, Yan Wang, Qi Dou, Xiaokang Yang, and Wei Shen. *EndoGSLAM: Real-Time Dense Reconstruction and Tracking in Endoscopic Surgeries using Gaussian Splatting*. 2024. arXiv: 2403.15124 [cs.CV]. URL: <https://arxiv.org/abs/2403.15124> (cit. on pp. 14, 32).
- [8] Ruibin Ma, Rui Wang, Yubo Zhang, Chen Chen, Yao Song, Peng Wang, and Pheng-Ann Heng. «RNNSLAM: Reconstructing the 3D colon to visualize missing regions during a colonoscopy». In: *Medical Image Analysis* (2021) (cit. on pp. 14, 29).
- [9] Richard Elvira, Juan D. Tardós, and José M. M. Montiel. *CudaSIFT-SLAM: multiple-map visual SLAM for full procedure mapping in real human endoscopy*. 2024 (cit. on p. 14).

- [10] Javier Morlana, Juan D. Tardós, and José M. M. Montiel. *Topological SLAM in colonoscopies leveraging deep features and topological priors*. 2024 (cit. on p. 14).
- [11] Zhuoyue Yang, Ju Dai, and Junjun Pan. «3D reconstruction from endoscopy images: A survey». In: *Computers in Biology and Medicine* (2024) (cit. on p. 14).
- [12] Mohammad Khademul Bashar, Kenji Kondo, Yasuo Nomura, and Hideto Okada. «Structure-from-Motion from an Uncalibrated Endoscope». In: *3D Research* (2011) (cit. on p. 15).
- [13] Ruo Zhang, Ping-Sing Tsai, J.E. Cryer, and M. Shah. «Shape-from-shading: a survey». In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (1999) (cit. on p. 16).
- [14] Javier Ruano, Danail Stoyanov, Tom Vercauteren, Jan Deprest, Sebastien Ourselin, and Emmanuel Vander Poorten. «Learning shape from shading for colonoscopy». In: *Computerized Medical Imaging and Graphics* (2024) (cit. on p. 16).
- [15] Jens Ackermann, Michael Goesele, et al. «A survey of photometric stereo techniques». In: *Foundations and Trends® in Computer Graphics and Vision* (2015) (cit. on p. 17).
- [16] 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». In: *Sensors* (2020) (cit. on p. 17).
- [17] Víctor M. Batlle, J.M.M. Montiel, and Juan D. Tardós. «Photometric single-view dense 3D reconstruction in endoscopy». In: *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2022 (cit. on p. 17).
- [18] Ke Niu, Zeyun Liu, Xue Feng, Heng Li, Qika Lin, and Kaize Shi. *Endoscopic Depth Estimation Based on Deep Learning: A Survey*. 2025 (cit. on p. 18).
- [19] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. «Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data». In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2024, pp. 10371–10381 (cit. on pp. 18, 52).
- [20] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. *Depth Anything V2*. 2024. arXiv: 2406.09414 [cs.CV]. URL: <https://arxiv.org/abs/2406.09414> (cit. on p. 18).

- [21] Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. *Video Depth Anything: Consistent Depth Estimation for Super-Long Videos*. 2025. arXiv: 2501.12375 [cs.CV]. URL: <https://arxiv.org/abs/2501.12375> (cit. on pp. 18, 53).
- [22] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. *Depth Pro: Sharp Monocular Metric Depth in Less Than a Second*. 2025. arXiv: 2410.02073 [cs.CV]. URL: <https://arxiv.org/abs/2410.02073> (cit. on pp. 18, 51).
- [23] Ziang Xu, Bin Li, Yang Hu, Chenyu Zhang, James East, Sharib Ali, and Jens Rittscher. *Self-supervised Monocular Depth and Pose Estimation for Endoscopy with Generative Latent Priors*. 2024. arXiv: 2411.17790 [cs.CV]. URL: <https://arxiv.org/abs/2411.17790> (cit. on pp. 18, 19, 52).
- [24] Romain Hardy, Tyler Berzin, and Pranav Rajpurkar. *ColonCrafter: A Depth Estimation Model for Colonoscopy Videos Using Diffusion Priors*. 2025. arXiv: 2509.13525 [cs.CV]. URL: <https://arxiv.org/abs/2509.13525> (cit. on pp. 18, 52).
- [25] X. Anadón, Javier Rodríguez-Puigvert, and J. M. M. Montiel. *3D Densification for Multi-Map Monocular VSLAM in Endoscopy*. 2025 (cit. on p. 19).
- [26] Anita Rau, Binod Bhattarai, Lourdes Agapito, and Danail Stoyanov. «Bi-modal Camera Pose Prediction for Endoscopy». In: *IEEE Transactions on Medical Robotics and Bionics* 5.4 (Nov. 2023), pp. 978–989. ISSN: 2576-3202. DOI: 10.1109/tmr.2023.3320267. URL: <http://dx.doi.org/10.1109/TMRB.2023.3320267> (cit. on pp. 19, 57, 59).
- [27] Ruyu Liu, Zhengzhe Liu, Haoyu Zhang, Guodao Zhang, Jianhua Zhang, Sunbo, Weiguo Sheng, Xiufeng Liu, and Yaochu Jin. «ColVO: Colonoscopic Visual Odometry Considering Geometric and Photometric Consistency». In: *Proceedings of the 32nd ACM International Conference on Multimedia*. MM ’24. Melbourne VIC, Australia: Association for Computing Machinery, 2024, pp. 8100–8109. ISBN: 9798400706868. DOI: 10.1145/3664647.3681286. URL: <https://doi.org/10.1145/3664647.3681286> (cit. on pp. 20, 29).
- [28] Ruyu Liu, Zhengzhe Liu, Haoyu Zhang, Guodao Zhang, Jianhua Zhang, Bo Sun, Weiguo Sheng, Xiufeng Liu, and Yaochu Jin. «Sparse-to-dense coarse-to-fine depth estimation for colonoscopy». In: *Medical Image Analysis* 91 (2024), p. 103025. DOI: 10.1016/j.media.2024.103025 (cit. on p. 23).
- [29] David Recasens, José Lamarca, José M. Fácil, J. M. M. Montiel, and Javier Civera. «Endo-Depth-and-Motion: Reconstruction and Tracking in Endoscopic Videos Using Depth Networks and Photometric Constraints». In: *IEEE Robotics and Automation Letters* 6.4 (2021), pp. 7225–7232. DOI: 10.1109/LRA.2021.3095528 (cit. on p. 23).

- [30] 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*. 2020. arXiv: 2003.08934 [cs.CV]. URL: <https://arxiv.org/abs/2003.08934> (cit. on p. 24).
- [31] Yufei Shi, Beijia Lu, Jia-Wei Liu, Ming Li, and Mike Zheng Shou. *ColonNeRF: High-Fidelity Neural Reconstruction of Long Colonoscopy*. 2024. arXiv: 2312.02015 [cs.CV]. URL: <https://arxiv.org/abs/2312.02015> (cit. on p. 24).
- [32] Yuehao Wang, Bingchen Gong, Yonghao Long, Siu Hin Fan, and Qi Dou. *Efficient EndoNeRF Reconstruction and Its Application for Data-driven Surgical Simulation*. 2024. arXiv: 2404.15339 [eess.IV]. URL: <https://arxiv.org/abs/2404.15339> (cit. on p. 24).
- [33] Lingting Zhu, Zhao Wang, Jiahao Cui, Zhenchao Jin, Guying Lin, and Lequan Yu. *EndoGS: Deformable Endoscopic Tissues Reconstruction with Gaussian Splatting*. 2024. arXiv: 2401.11535 [cs.CV]. URL: <https://arxiv.org/abs/2401.11535> (cit. on p. 26).
- [34] Yifan Liu, Chenxin Li, Chen Yang, and Yixuan Yuan. *EndoGaussian: Real-time Gaussian Splatting for Dynamic Endoscopic Scene Reconstruction*. 2024. arXiv: 2401.12561 [cs.CV]. URL: <https://arxiv.org/abs/2401.12561> (cit. on p. 26).
- [35] Yiming Huang, Beilei Cui, Long Bai, Zhen Chen, Jinlin Wu, Zhen Li, Hongbin Liu, and Hongliang Ren. *Advancing Dense Endoscopic Reconstruction with Gaussian Splatting-driven Surface Normal-aware Tracking and Mapping*. 2025. arXiv: 2501.19319 [cs.CV]. URL: <https://arxiv.org/abs/2501.19319> (cit. on pp. 26, 31).
- [36] Perry J Pickhardt, Andrew J Taylor, and Deepak K Gopal. «Surface Visualization at 3D Endoluminal CT Colonography: Degree of Coverage and Implications for Polyp Detection». In: *Gastroenterology* 129.6 (2005). DOI: 10.1053/j.gastro.2005.09.008 (cit. on p. 28).
- [37] Ruibin 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: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part V*. Shenzhen, China: Springer-Verlag, 2019. ISBN: 978-3-030-32253-3. DOI: 10.1007/978-3-030-32254-0_64. URL: https://doi.org/10.1007/978-3-030-32254-0_64 (cit. on p. 28).

- [38] Netanel Frank, Erez Posner, Emmanuelle Muhlethaler, Adi Zholkover, and Moshe Bouhnik. *ColNav: Real-Time Colon Navigation for Colonoscopy*. 2023. arXiv: 2306.04269 [cs.CV]. URL: <https://arxiv.org/abs/2306.04269> (cit. on p. 28).
- [39] Emmanuelle Muhlethaler, Erez Posner, and Moshe Bouhnik. *Estimating the coverage in 3d reconstructions of the colon from colonoscopy videos*. 2022. arXiv: 2210.10459 [cs.CV]. URL: <https://arxiv.org/abs/2210.10459> (cit. on p. 28).
- [40] Ruibin 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: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Ed. by Dinggang Shen, Tianming Liu, Terry M. Peters, et al. Springer International Publishing, 2019. DOI: 10.1007/978-3-030-32254-0_64 (cit. on p. 29).
- [41] Erez Posner, Adi Zholkover, Netanel Frank, and Moshe Bouhnik. *C³Fusion: Consistent Contrastive Colon Fusion, Towards Deep SLAM in Colonoscopy*. 2022. arXiv: 2206.01961 [cs.CV]. URL: <https://arxiv.org/abs/2206.01961> (cit. on p. 32).
- [42] Beilei Cui et al. *Learning to Efficiently Adapt Foundation Models for Self-Supervised Endoscopic 3D Scene Reconstruction from Any Cameras*. 2025. arXiv: 2503.15917 [cs.CV]. URL: <https://arxiv.org/abs/2503.15917> (cit. on p. 39).
- [43] Nicolas Toussaint, Emanuele Colleoni, Ricardo Sanchez-Matilla, Joshua Sutcliffe, Vanessa Thompson, Muhammad Asad, Imanol Luengo, and Danail Stoyanov. «Zero-Shot Monocular Metric Depth for Endoscopic Images». In: *MICCAI Workshop on Data Engineering in Medical Imaging*. Springer. 2025, pp. 115–124 (cit. on p. 52).
- [44] Zanwei Zhou, Chen Yang, Piao Yang, Xiaokang Yang, and Wei Shen. «EndoDAV: Depth Any Video in Endoscopy with Spatiotemporal Accuracy». In: *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*. Vol. LNCS 15968. Springer Nature Switzerland, Sept. 2025 (cit. on p. 53).