Deep learning-based Image exposure correction and depth estimation for 3D cartography in endoscopy

Ricardo Espinosa

Deep learning-based Image exposure correction and depth estimation for 3D cartography in endoscopy

THÈSE

présentée et soutenue publiquement le 20 novembre 2025

pour l'obtention du

Doctorat de l'Université de Lorraine

Spécialité Automatique, Traitement du Signal et des Images, Génie Informatique

par

Ricardo Abel Espinosa Loera

Composition du jury

Président :	**Marie-Odile Berger**	Directrice de Recherche INRIA, LORIA.
Rapporteurs :	**David Fofi**	PU, Université de Bourgogne, Laboratoire ImViA.
	Jean-Bernard Hayet	CIMAT, Guanajuato, Mexique.
Examinatrices :	**Mariel Alfaro-Ponce**	Tecnológico de Monterrey, campus Mexico City
	Marie-Odile Berger	Directrice de Recherche INRIA, LORIA.
Co-directeurs :	**Christian Daul**	PU, Université de Lorraine, CRAN.
	Gilberto Ochoa-Ruiz	Tecnológico de Monterrey, campus Guadalajara.

Centre de Recherche en Automatique de Nancy
UMR 7039 - Université de Lorraine, CNRS

Acknowledgments ◂

I take this opportunity to acknowledge those who have been part of this important step in my career. First, I would like to acknowledge my family — Samantha, Sarah, and Ricardo — for supporting me and suffering with me through this long journey called a PhD. You have always been by my side, and I want you to know that all of this effort has been for you. Thank you so much — I love you.

To my parents and brothers, thank you for your words of support always, but most importantly, for loving me unconditionally.

To my advisors, I truly believe I have been very lucky to be taught and guided by top researchers and professors — Prof. Christian and Prof. Gilberto. But most importantly, they are genuinely good people: kind yet strict, always willing to help me when I struggled, and constantly looking for ways to help me grow. Thank you so much for your patience and guidance throughout this long process.

To my institutions: Universidad Panamericana, thank you for supporting me in so many ways since the beginning of this journey. Thank you for your openness and for letting me grow abroad. Also, thank you to UL for helping me discover a better version of myself.

Contents

Abstract / Résumé

English version:

This thesis investigates the application of deep learning techniques to improve 3D reconstruction in endoscopy. A passive vision framework is proposed to enhance monocular video data by means of photometric correction and self-supervised depth estimation, addressing challenges related to uneven illumination and limited geometric information. The contributions include the development of *Endo4IE*, a synthetic dataset specifically designed to support the training of exposure correction models, as well as the implementation of deep-learning methods capable of recovering structural details in poorly illuminated regions. In addition, an illumination-invariant and self-supervised depth estimation model is introduced to infer scene geometry from monocular sequences without requiring ground-truths. These modules have been integrated into a 3D reconstruction pipeline enabling the generation of dense surfaces with coherent shapes. Extensive experiments on both synthetic and real colonoscopic data demonstrate significant improvements in i) the image contrasts, ii) the accuracy of the 3D camera trajectory tracking, and iii) the quality of the surface construction. These results all contribute to a more reliable and geometry-aware visualization in clinical endoscopy.

Version française :

Ce manuscrit porte sur le développement de techniques d'apprentissage profond pour améliorer la reconstruction 3D en endoscopie. Ce travail propose, dans le cadre de la construction de surfaces à partir de vidéos-séquences 2D d'endoscopie monoculaire, une approche de vision passive qui combine une correction photométrique et une estimation de profondeur auto-supervisée, le défi étant de prendre en compte les difficultés liées à l'illumination non uniforme des parois internes du côlon et à la présence de peu de textures et de structure dans les images. Les contributions incluent la création du jeu de données synthétique *Endo4IE* conçu pour l'entraînement de modèles de correction de l'exposition, ainsi que de l'amélioration de méthodes d'apprentissage profond capables de restaurer les détails structurels dans les zones d'images sous- ou surexposées. Par ailleurs, un modèle d'estimation de profondeur auto-supervisé et invariant aux changements de l'illumination a été proposé pour calculer la forme des surfaces sans vérité terrain. Ces modules ont été intégrés dans une chaîne de reconstruction 3D permettant la génération de surfaces denses aux formes cohérentes. De nombreuses expériences avec des données synthétiques et réelles démontrent l'amélioration i) des contrastes dans les images, ii) de la robustesse du suivi de la trajectoire 3D de la caméra dans le côlon et iii) de la qualité de la construction des surfaces. Ces résultats contribuent tous à une visualisation endoscopique plus exploitable et guidée par la géométrie.

Introduction Générale

Les travaux décrits dans ce manuscrit ont été réalisés dans le cadre d'un programme doctoral en cotutelle entre l'Universidad Panamericana (Mexique) et l'Université de Lorraine (Centre de Recherche en Automatique de Nancy, UMR 7039, CNRS/UL), France.

Le contexte de cette thèse se situe à l'intersection de l'imagerie médicale, de la vision par ordinateur et de l'apprentissage profond, l'objectif étant de développer des méthodes de calcul visant à améliorer la précision et l'utilisabilité des données endoscopiques pour des finalités diagnostiques et interventionnelles. Ce travail a été motivé par un intérêt partagé entre cliniciens et ingénieurs : surmonter les limitations technologiques qui entravent l'exploration complète et la documentation exhaustive de la surface muqueuse interne du côlon.

Cette recherche s'attaque aux limitations critiques de la coloscopie conventionnelle en intégrant un traitement d'images fondé sur l'apprentissage profond dans des chaînes de reconstruction 3D. La coloscopie, bien qu'elle soit considérée comme la procédure de référence pour la détection et l'ablation de polypes précancéreux et de carcinomes colorectaux à un stade précoce, souffre de limitations significatives en termes de couverture spatiale complète de la paroi interne du côlon ainsi que pour l'interprétation de la scène à des fins diagnostiques. Ces limitations proviennent du champ de vision intrinsèquement étroit des endoscopes, des conditions d'illumination hautement variables, des réflexions spéculaires et du caractère fondamentalement bidimensionnel des séquences vidéo acquises. Ensemble, ces facteurs réduisent l'interprétabilité des images, peuvent conduire à des lésions manquées et compliquent le suivi longitudinal des patients.

Dans la pratique clinique actuelle, l'évaluation des données endoscopiques affichées sur un écran repose souvent sur une inspection visuelle en temps réel pendant la procédure (ici une coloscopie), avec un soutien limité pour l'analyse rétrospective ou l'intégration dans les dossiers numériques. Bien que diverses techniques de mosaïquage 2D aient été proposées pour étendre le champ de vision, elles introduisent fréquemment des distorsions géométriques, une dégradation de la résolution et échouent à préserver la véritable profondeur spatiale, laquelle constitue un facteur essentiel pour la compréhension de formes anatomiques complexes et pour le recalage de données entre différentes sessions ou modalités. L'absence de représentations 3D cohérentes limite non seulement la confiance diagnostique, mais entrave également l'adoption d'outils d'assistance computationnelle tels que les systèmes de navigation, la détection assistée par IA ou le suivi longitudinal des lésions.

À l'inverse, les techniques de reconstruction 3D — en particulier celles basées sur le *Structure-from-Motion* (SfM) — présentent un potentiel important pour construire des modèles étendus et géométriquement précis de la muqueuse colique. Toutefois, l'application du SfM classique en endoscopie s'avère extrêmement difficile. Les images coloscopiques manquent souvent de textures distinctives, souffrent de flous de mouvement et d'occlusions, et sont acquises dans des conditions d'illumination non contrôlées. Ces facteurs dégradent la fiabilité de la détection de caractéristiques, de l'appariement de points correspondants et de l'estimation de la pose — éléments fondamentaux de toute chaîne de reconstruction géométrique. Des techniques SfM récentes ont été adaptées pour la reconstruction 3D de la paroi gastrique (scènes peu texturées et présentant des changements d'illumination importants entre les points de vue). Cependant, ces approches sont computationnellement intensives et ne peuvent être utilisées qu'en mode différé (par exemple une heure après l'examen). En coloscopie, une reconstruction en temps réel est requise pour visualiser les parties de surface qui n'ont pas été explorées par l'endoscope.

Cette thèse introduit un cadre de vision passive novateur intégrant une amélioration d'image fondée sur l'apprentissage profond et une estimation auto-supervisée de profondeur directement dans des processus SfM ou SLAM (*Simultaneous Localization and Mapping*). Contrairement aux méthodes actives, qui nécessitent des modifications matérielles (par exemple l'ajout de lumière structurée), l'approche de vision passive proposée fonctionne exclusivement sur des vidéos coloscopiques monoculaires standards. Cela garantit une compatibilité avec les équipements cliniques actuels tout en offrant des améliorations significatives en termes de qualité d'image et de cohérence spatiale. Les contributions principales de cette thèse incluent :

— le développement d'Endo4IE, un jeu de données synthétique apparié, spécifiquement conçu pour soutenir l'entraînement et l'évaluation de méthodes d'amélioration photométrique fondées sur l'apprentissage profond dans l'imagerie endoscopique ;

— la conception de méthodes d'ajustement photométrique capables d'atténuer la sous-exposition et la surexposition locales apparaissant dans différentes régions de l'image en fonction de la géométrie interne du côlon. L'atténuation de ces expositions non optimales permet de récupérer des détails structuraux critiques et d'améliorer la cohérence inter-images ;

— le développement d'un cadre auto-supervisé d'estimation de profondeur monoculaire invariant à l'illumination, exploitant des contraintes géométriques et photométriques sur des séquences vidéo pour apprendre la profondeur sans annotations de vérité terrain ;

— l'intégration de ces modules dans une méthode de construction de surfaces 3D permettant la reconstruction de modèles de surface denses, haute résolution et topologiquement cohérents du côlon à partir de vidéos monoculaires ;

— une évaluation expérimentale complète sur des jeux de données synthétiques et des séquences coloscopiques réelles, mesurant les performances en termes de qualité d'image, de précision

de suivi caméra et de complétude de reconstruction.

À travers ces contributions, cette thèse vise à réduire l'écart entre l'amélioration photométrique de bas niveau et la modélisation géométrique de haut niveau dans le contexte endoscopique. Le cadre proposé offre une solution évolutive et potentiellement applicable en pratique clinique pour générer des représentations 3D étendues fidèles à l'anatomie. Ce faisant, il ouvre la voie à un diagnostic assisté par ordinateur plus robuste, à une meilleure documentation et localisation des lésions, ainsi qu'à une continuité de soin améliorée grâce à des enregistrements numériques enrichis et spatialement cohérents des procédures endoscopiques.

Chapitre 1 : Contexte médical et objectifs scientifiques.

Ce chapitre introduit le contexte clinique et technologique qui motive cette thèse. Il commence par décrire le rôle de l'endoscopie dans le diagnostic gastro-intestinal moderne, en mettant particulièrement l'accent sur les défis liés à la visualisation de la surface muqueuse interne du côlon. Ces défis découlent du champ de vision restreint, des conditions d'éclairage dynamiques et de la géométrie complexe de la paroi colique. Dans ce contexte, le chapitre souligne l'importance médicale de construire des représentations 3D étendues et géométriquement cohérentes à partir de séquences vidéo coloscopiques standards.

Le chapitre poursuit avec une revue des principales classes de techniques de reconstruction 3D applicables à l'endoscopie, notamment les méthodes de vision active et passive. Chaque approche est discutée selon sa pertinence clinique, ses exigences matérielles et sa robustesse face aux contraintes caractéristiques de l'imagerie endoscopique — telles que le manque de texture, les reflets spéculaires et les variations de mouvement de la caméra. Sur la base de cette analyse, l'étude argue en faveur de l'utilisation de la vision passive et plus spécifiquement du *Structure-from-Motion* (SfM) comme fondement pour la cartographie 3D en coloscopie.

Une présentation détaillée du cadre SfM est ensuite fournie, incluant ses composants algorithmiques essentiels et les architectures de pipeline. Les défis liés à l'application des techniques SfM classiques aux données endoscopiques sont identifiés et discutés, notamment la difficulté d'extraire des correspondances fiables et l'impact de l'incohérence photométrique sur l'estimation de pose. Le chapitre examine également les trois principales catégories de pipelines SfM — globaux, hiérarchiques et incrémentaux — et justifie le choix d'un pipeline SfM incrémental comme solution la plus adaptée pour traiter des séquences coloscopiques monoculaires.

En conclusion, ce chapitre pose les bases du reste du manuscrit en identifiant les défis scientifiques fondamentaux associés au développement d'une chaîne de reconstruction 3D robuste et cliniquement applicable en coloscopie. Il introduit l'hypothèse centrale selon laquelle l'intégration d'un prétraitement photométrique fondé sur l'apprentissage profond avec des techniques SLAM incrémentales permet de surmonter les limitations des méthodes conventionnelles et d'obtenir des reconstructions 3D de haute fidélité à partir de vidéos coloscopiques monoculaires en conditions

cliniques réelles.

Chapitre 2 : Amélioration d'image par apprentissage profond en endoscopie.

Ce chapitre introduit le problème de la dégradation d'image en coloscopie due à la sous-exposition, la surexposition et les réflexions spéculaires, phénomènes fréquents et problématiques en conditions cliniques réelles. Ces effets altèrent l'interprétation visuelle des images et compromettent fortement la performance des algorithmes de reconstruction 3D. Le chapitre motive la nécessité de techniques d'amélioration dédiées capables de restaurer les détails structuraux, de corriger les incohérences photométriques et d'homogénéiser la qualité d'image au sein des séquences vidéo.

Le chapitre débute par une vue d'ensemble des méthodes classiques d'amélioration d'image, incluant les approches basées sur l'histogramme, les modèles d'illumination et les corrections inspirées de la théorie de Retinex. Ces méthodes traditionnelles, bien que rapides et interprétables, se révèlent souvent insuffisantes face aux artefacts photométriques localisés et aux gradients d'illumination abrupts rencontrés dans les images endoscopiques.

Pour dépasser ces limitations, le chapitre présente un ensemble de modèles d'amélioration développés pour les images coloscopiques : Endo-LMSPEC, un réseau multiscalaire fondé sur la pyramide laplacienne conçu pour restaurer le contraste local et global, et Endo-ViT, une architecture Transformer exploitant des mécanismes d'attention et des contraintes sur les distributions de couleurs pour produire des corrections photométriques cohérentes et respectueuses des détails. Un modèle d'inpainting temporel (Endo-STTN) est également introduit pour traiter les reflets spéculaires en reconstruisant les régions saturées à partir de la cohérence inter-images.

Le développement de ces modèles repose sur le jeu de données synthétique Endo-4IE, créé dans cette thèse pour fournir des paires d'images dégradées et de référence sous conditions contrôlées. Le pipeline de synthèse d'image est étudié en détail, de même que les protocoles d'évaluation quantitatifs et qualitatifs utilisés pour juger les performances des modèles.

La dernière section du chapitre présente une évaluation complète des méthodes proposées, incluant des résultats visuels, des mesures quantitatives (SSIM, PSNR, NIQE) et des expériences d'ablation. L'intégration de ces modules comme étapes de prétraitement dans une chaîne de reconstruction 3D est enfin introduite, préparant le terrain pour les expérimentations du Chapitre 3. Dans l'ensemble, ce chapitre démontre que l'amélioration fondée sur l'apprentissage profond améliore considérablement la qualité d'image et fournit une entrée plus stable pour les algorithmes de reconstruction géométrique en coloscopie.

Chapitre 3 : Intégration des modules d'amélioration dans une chaîne de reconstruction 3D.

Ce chapitre présente l'intégration des modèles d'amélioration d'image développés dans le Chapitre 2 au sein d'un pipeline SLAM monoculaire dédié à la reconstruction 3D en coloscopie. L'objectif est d'évaluer comment les améliorations photométriques obtenues via l'apprentissage profond influencent la précision de l'estimation de trajectoire et la complétude des surfaces reconstruites.

Le chapitre décrit le pipeline de reconstruction utilisé, basé sur une version modifiée de RNN-SLAM, adaptée aux vidéos coloscopiques. Celui-ci intègre une estimation de pose image-à-image et une reconstruction incrémentale de surface, tout en permettant l'injection de cadres améliorés avant la prédiction de profondeur et l'étape de fusion.

Trois méthodes d'amélioration sont testées : Endo-LMSPEC pour la correction d'exposition multiscalaire, Endo-STTN pour l'inpainting spéculaire, et une correction gamma récurrente. Chacune est appliquée comme prétraitement, et son impact est évalué en termes d'estimation de trajectoire et de complétude des maillages reconstruits.

Les évaluations sont conduites sur des jeux de données synthétiques et réelles. Des métriques classiques de trajectoire (ATE, RPE) ainsi que des comparaisons qualitatives des surfaces reconstruites sont fournies. Les résultats expérimentaux montrent que l'amélioration d'image renforce la stabilité et la précision du pipeline SLAM dans des conditions d'illumination difficiles, réduisant la dérive et améliorant la cohérence des prédictions de profondeur.

Ce chapitre démontre que l'intégration directe de modules d'amélioration fondés sur l'apprentissage profond améliore la robustesse et la qualité de la cartographie de surface 3D monoculaire en coloscopie.

Chapitre 4 : Estimation de profondeur auto-supervisée en endoscopie.

Ce chapitre introduit un cadre innovant d'apprentissage auto-supervisé pour l'estimation de profondeur monoculaire adapté aux contraintes spécifiques de l'imagerie endoscopique. La profondeur joue un rôle fondamental dans les pipelines de reconstruction 3D traditionnels, car elle permet de reconstruire des surfaces anatomiques cohérentes. Toutefois, en coloscopie, l'absence de vérités terrain pour la profondeur rend l'apprentissage supervisé impraticable, et les méthodes classiques peinent à faire face aux variations photométriques, au manque de texture et à la non-rigidité des tissus.

Pour répondre à ces limitations, ce chapitre introduit une stratégie d'apprentissage auto-supervisé exploitant les relations géométriques entre images consécutives sans nécessiter d'annotations de profondeur. L'approche proposée repose sur une architecture encodeur-décodeur à base de Transformers, capable de modéliser les dépendances globales et adaptée aux environnements endoscopiques faiblement texturés. En outre, le chapitre introduit un objectif d'apprentissage invariant à l'illumination, intégrant une perte photométrique sensible aux occultations, un terme de cohérence structurelle et des pénalités locales préservant le contraste.

Deux modèles sont développés et analysés : MonoII, un modèle léger assurant une base solide, et MonoIIT, une variante plus élaborée intégrant des capacités de raisonnement global accrues grâce aux Transformers. Les évaluations quantitatives et qualitatives sont effectuées sur des données synthétiques et réelles, démontrant la capacité de la méthode à produire des cartes de profondeur plausibles sous de fortes variations d'illumination.

Les résultats montrent que la stratégie auto-supervisée constitue une solution viable et pertinente pour la coloscopie monoculaire, fournissant des cartes de profondeur essentielles à la reconstruction 3D.

Chapitre 5 : Reconstruction 3D de différents organes internes.

Ce chapitre étend la méthodologie développée au-delà de la coloscopie afin de démontrer sa généralisabilité à divers organes et contextes endoscopiques. Il examine l'applicabilité du pipeline complet — amélioration d'image, estimation de profondeur auto-supervisée et SLAM photométrique — dans des environnements endoscopiques et chirurgicaux variés, incluant des données ex-vivo et in-vivo.

Les résultats montrent que la méthode proposée reconstruit avec succès des surfaces étendues malgré des conditions d'illumination variées et la déformabilité importante des organes. Pour l'estomac et l'intestin grêle, les surfaces reconstruites présentent une cohérence topologique convaincante et une bonne restitution des plis muqueux. Pour les séquences coloscopiques in vivo, les reconstructions préservent les structures essentielles nécessaires à une interprétation clinique fiable.

Le chapitre discute également des limitations observées, notamment dans les régions extrêmement pauvres en structure ou fortement affectées par les artefacts de mouvement, ouvrant la voie à des travaux futurs sur la modélisation des déformations.

En résumé, ce chapitre démontre que l'approche proposée constitue un cadre général pour la reconstruction 3D monoculaire dans différents organes internes, avec un potentiel significatif pour la navigation médicale et la planification chirurgicale.

Conclusion générale.

La conclusion générale résume les contributions majeures de cette thèse, notamment le développement du jeu de données Endo4IE, la proposition de méthodes d'amélioration d'image basées sur les CNNs et Transformers, l'introduction d'un cadre d'estimation de profondeur auto-supervisée invariant à l'illumination, et l'intégration de ces modules dans une chaîne de reconstruction 3D cohérente. Des perspectives de recherche sont proposées, incluant l'intégration de représentations implicites telles que les NeRFs, l'amélioration du traitement en temps réel, et l'extension de la méthode à des environnements chirurgicaux plus complexes.

General Introduction

The works described in the manuscript were carried out in the framework of a joint doctoral program between the Universidad Panamericana (Mexico) and the Université de Lorraine (Centre de Recherche en Automatique de Nancy, UMR 7039, CNRS/UL), France.

The thesis context is situated at the intersection of medical imaging, computer vision, and deep learning, the objective being to develop computational methods that improve the accuracy and usability of endoscopic data for diagnostic and interventional purposes. The work has been driven by a shared motivation between clinicians and engineers: to overcome the technological limitations that hinder the comprehensive exploration and documentation of the internal mucosal surface of the colon.

This research addresses critical limitations in conventional colonoscopy through the integration of deep learning-based image processing into 3D reconstruction pipelines. Colonoscopy, while regarded as the gold-standard procedure for the detection and removal of precancerous polyps and early-stage colorectal carcinomas, suffers from significant limitations in terms of a complete spatial coverage of the inner colon wall scene interpretation for diagnosis. These limitations stem from the inherently narrow field of view of endoscopes, highly variable illumination conditions, specular reflections, and the fundamentally two-dimensional nature of the acquired video sequences. Together, these factors reduce the interpretability of the images, can lead to missed lesions, and complicate longitudinal follow-up of patients.

In current clinical practice, the evaluation of endoscopic data displayed on a screen often relies on real-time visual inspection during the procedure (here a colonoscopy), with limited support for retrospective analysis or integration into digital records. Although various 2D mosaicing techniques have been proposed to extend the field of view, they frequently introduce geometric distortions, resolution degradation, and fail to preserve true spatial depth which is an essential factor for understanding complex anatomical shapes and for registering data across sessions or modalities. The lack of consistent 3D representations not only limits diagnostic confidence but also impedes the adoption of computer-assisted tools such as navigation systems, AI-assisted detection, or longitudinal lesion tracking.

In contrast, 3D reconstruction techniques, particularly those based on Structure-from-Motion (SfM), hold great potential for constructing extended, geometrically accurate models of the colonic

mucosa. However, the application of classical SfM in endoscopy has proven extremely challenging. Colonoscopic images often lack strong textures, suffer from motion blur and occlusions, and are acquired under uncontrolled lighting conditions. These factors degrade the reliability of feature detection, matching, and pose estimation—core components of any geometry-based reconstruction pipeline. Recent SfM techniques were adapted to the 3D reconstruction of the stomach wall (scenes with few textures and complex illumination changes between viewpoints). However, these approaches are computation intensive and are only usable in offline (an hour after the examination). In colonoscopy, a real time reconstruction is required to visualize surface parts that were not scanned by the endoscope.

This thesis introduces a novel passive vision framework that integrates deep learning-based image enhancement and self-supervised depth estimation directly into SfM or SLAM (Simultaneous Localization And Mapping) processes. Unlike active vision methods, which require hardware modifications based on structured light for instance, the proposed passive vision approach operates purely on standard monocular colonoscopic videos. This ensures compatibility with current clinical hardware and workflows while offering significant improvements in image quality and spatial coherence. The key contributions of the thesis include:

— The development of Endo4IE, a synthetic paired dataset specifically designed to support the training and evaluation of deep learning-based photometric enhancement methods in endoscopic imaging.

— The design of learning-based photometric correction methods capable of attenuating local underexposure and overexposure often arising in various image regions according to the inner colon shape. The attenuation of non optimal exposure allows to recover critical structural details and enhancing inter-frame consistency;

— The development of an illumination-invariant, self-supervised monocular depth estimation framework that leverages geometric and photometric constraints across sequences to learn depth without requiring ground truth annotations;

— The integration of these modules into a 3D surface construction method enabling the reconstruction of dense, high-resolution, and topologically consistent 3D surface models of the colon from monocular video;

— A comprehensive experimental evaluation using both synthetic datasets and real-world colonoscopic sequences, measuring performance in terms of image quality, camera tracking accuracy, and reconstruction completeness.

Through these contributions, the thesis aims to bridge the gap between low-level photometric enhancement and high-level geometric modeling in the endoscopic context. The proposed framework offers a scalable, and potentially clinically applicable solution for generating extended 3D field-of-view representations that are anatomically faithful. In doing so, it paves the way for more

robust computer-aided diagnosis, improved lesion documentation and localization, and enhanced continuity of care through standardized, spatially rich digital records of endoscopic procedures.

Chapter 1: Medical context and scientific objectives.

This chapter introduces the clinical and technological background that motivates this thesis. It begins by describing the role of endoscopy in modern gastrointestinal diagnosis, with a particular focus on the challenges of visualizing the internal mucosa surface of the colon. These challenges arise from the narrow field of view, dynamic lighting conditions, and the complex geometry of the colonic wall. In this context, the chapter emphasizes the medical relevance of constructing extended and geometrically consistent 3D surface representations from standard colonoscopic video sequences. The chapter proceeds with a review of the major classes of 3D reconstruction techniques applicable to endoscopy, namely active and passive vision methods. Each approach is discussed with respect to its clinical applicability, hardware requirements, and robustness under the constraints typical of endoscopic imaging—such as low texture, specular highlights, and camera motion variability. Based on this analysis, the study argues for the suitability of passive vision strategies, and more specifically, for the use of Structure-from-Motion (SfM) as a foundation for 3D cartography in colonoscopy. A detailed presentation of the SfM framework is then provided, including its core algorithmic components and pipeline architectures. The challenges of applying classical SfM techniques to endoscopic data are identified and discussed, such as the difficulty of extracting reliable feature correspondences and the impact of photometric inconsistency on camera pose estimation. The chapter also reviews the three principal categories of SfM pipelines—global, hierarchical, and incremental—and justifies the choice of an incremental SfM pipeline as the most appropriate for handling monocular colonoscopy sequences. In conclusion, this chapter lays the groundwork for the remainder of the thesis by identifying the core scientific challenges involved in developing a robust and clinically applicable 3D reconstruction pipeline for colonoscopy. It introduces the central hypothesis that the integration of deep learning-based image preprocessing with incremental SLAM techniques can effectively overcome the limitations of conventional methods, thereby enabling high-fidelity 3D surface reconstruction from monocular colonoscopic video in real-world clinical settings.

Chapter 2: Deep learning-based image enhancement in endoscopy.

This chapter introduces the problem of image degradation in colonoscopy due to underexposure, overexposure, and specular reflections, which are frequent and problematic in real clinical conditions. These issues affect the visual interpretability of the images and severely compromise the performance of downstream algorithms for 3D reconstruction. The chapter motivates the need for dedicated enhancement techniques that can restore structural details, correct photometric inconsistencies, and standardize image quality across video sequences. Chapter 2 begins with an overview of classical image enhancement methods, including histogram-based approaches, illumination mo-

deling techniques, and Retinex-inspired corrections. These traditional methods are computationally efficient and interpretable but often insufficient in the presence of localized photometric artifacts and strong illumination gradients, as commonly encountered in endoscopic imagery. To overcome these limitations, the chapter then presents a set of learning-based enhancement models specifically developed for colonoscopic images. These include Endo-LMSPEC, a multiscale Laplacian-based enhancement network trained to recover local and global contrast, and Endo-ViT, a transformer-based architecture that exploits attention mechanisms and histogram priors to achieve photometrically consistent and detail-preserving corrections. In addition, a temporal inpainting model (Endo-STTN) is introduced to address specular highlights and restore missing surface content using cross-frame consistency. The development and training of these models rely on the Endo-4IE dataset, a synthetic benchmark created to provide realistic pairs of degraded and reference-quality images under controlled conditions. The dataset is described in detail, along with the image synthesis pipeline used to simulate clinically relevant exposure artifacts. Evaluation metrics and experimental protocols are defined to quantitatively assess the performance of the proposed models. The final section of the chapter presents a comprehensive evaluation of the enhancement methods, including visual examples, quantitative metrics (e.g., SSIM, PSNR, NIQE), and ablation studies. The integration of these enhancement modules as preprocessing steps in a 3D reconstruction pipeline is also introduced, laying the foundation for the experiments described in Chapter 3. Overall, Chapter 2 demonstrates that learning-based enhancement significantly improves image quality and provides a stable input for geometry-based reconstruction algorithms in colonoscopy.

Chapter 3: Integration of DL-based image enhancement in 3D reconstruction pipelines.

This chapter presents the integration of the deep learning-based image enhancement models, developed in Chapter 2, into a monocular SLAM pipeline for 3D reconstruction in colonoscopy. The objective of this chapter is to evaluate how photometric improvements achieved through learning-based preprocessing influence the accuracy of camera trajectory estimation and the completeness of the reconstructed 3D surfaces. The chapter begins by describing the baseline reconstruction framework employed for experimentation. This pipeline is based on a modified version of RNN-SLAM, adapted for monocular colonoscopic video. It incorporates frame-to-frame pose estimation and incremental surface reconstruction, and is designed to accept enhanced frames as input prior to depth prediction and fusion. Three enhancement methods are individually tested in this setting: the proposed Endo-LMSPEC multiscale exposure correction network, the Endo-STTN transformer for specular inpainting, and a recurrent gamma correction module. Each module is applied as a preprocessing step, and its contribution is assessed in terms of its effect on trajectory estimation and 3D mesh completeness. Evaluation is conducted on both synthetic and real colonoscopy datasets. Quantitative performance is measured using standard trajectory metrics, including Absolute Trajectory Error (ATE) and Relative Pose Error (RPE), with COLMAP-based reconstructions used as ground truth. Qualitative assessments of the reconstructed surfaces are also provided to illustrate

improvements in continuity and coverage. Experimental results indicate that image enhancement improves the stability and precision of the SLAM pipeline under difficult visual conditions, such as variable lighting, overexposed regions, and specular artifacts. Enhanced inputs lead to better depth prediction consistency and reduced pose estimation drift. This chapter demonstrates that incorporating deep learning-based enhancement models directly into the reconstruction process improves the robustness and quality of monocular 3D surface mapping in colonoscopy.

Chapter 4: Self-supervised depth prediction in endoscopy.

This chapter introduces a novel self-supervised learning framework for monocular depth prediction tailored to the specific challenges of endoscopic imaging. In traditional 3D reconstruction pipelines, accurate depth maps are critical to reconstructing anatomically consistent surfaces. However, in colonoscopy, supervised training of depth networks is infeasible due to the lack of ground truth depth annotations, and classical methods struggle under the photometric variability, texture scarcity, and non-rigid anatomical structures typical of real clinical scenes. To overcome these limitations, this chapter presents a self-supervised depth estimation strategy that exploits geometric relationships between consecutive frames without requiring depth labels. The proposed approach is built around a transformer-based encoder-decoder architecture that models long-range dependencies and is well-suited for low-texture and spatially complex endoscopic environments. In addition to the architecture, the chapter introduces an illumination-invariant training objective that incorporates occlusion-aware photometric loss, a structural consistency term, and local contrast-preserving penalties to improve prediction robustness in the presence of lighting fluctuations and specularities. The chapter details the training strategy used on synthetic and real-world colonoscopic datasets, with specific attention to temporal frame selection, view synthesis constraints, and pose supervision through pre-trained SLAM trajectories. Two models are developed and analyzed: MonoII (a lightweight baseline) and MonoIIT (a transformer-based high-capacity model with enhanced global reasoning capabilities). Quantitative and qualitative evaluations are performed on synthetic benchmarks with known depth, as well as on real colonoscopic sequences. The results demonstrate that the proposed self-supervised method achieves competitive performance and generates geometrically plausible depth maps, even under strong photometric variation. Ablation studies further confirm the contribution of each loss component to the model's overall robustness. In summary, this chapter establishes self-supervised depth prediction as a viable and clinically relevant solution for monocular colonoscopy. The resulting depth maps serve as an essential input to the subsequent 3D reconstruction stage and form a cornerstone of the passive vision pipeline proposed in this thesis.

Chapter 5: 3D reconstruction of different internal organs.

This chapter extends the proposed 3D reconstruction framework beyond colonoscopy to demonstrate its generalizability across various organs and examinations. It investigates the applicability of the developed pipeline (which combines image enhancement, self-supervised depth esti-

mation, and SLAM-based surface mapping) in different endoscopic and surgical contexts, involving both ex-vivo and in-vivo data. Chapter 5 begins with an overview of the experimental setup and datasets used. These include porcine data acquired from the small intestine and stomach under controlled laboratory conditions, as well as in-vivo colonoscopic sequences from human patients. Each dataset is processed using the same pipeline described in previous chapters, without altering core parameters or requiring specific hardware calibration. This validates the robustness of the method under varying anatomical, textural, and lighting conditions. For each organ, the proposed MonoIIT model is used to predict dense depth maps, which are then fused into global 3D surfaces using volumetric integration techniques. Camera tracking is performed using photometric SLAM, and alignment quality is assessed both qualitatively—via visual inspection of mesh continuity and texture consistency—and quantitatively, using error metrics where ground truth is available. Results show that the proposed method successfully reconstructs extended surfaces in organs with highly deformable shapes and varying illumination. In the case of the stomach and small intestine, the pipeline adapts to wide luminal variations and produces topologically coherent meshes. For in-vivo colon sequences, reconstructed surfaces preserve structural features such as folds and vascular patterns, which are crucial for lesion localization and clinical interpretation. Chapter 5 also discusses the limitations observed when applying the method to regions with minimal visual structure or heavy motion artifacts. These findings motivate future work on deformation modeling and refinement techniques, which could enhance reconstruction fidelity in highly dynamic or non-rigid scenes. In summary, Chapter 5 demonstrates that the proposed pipeline is not restricted to colonoscopy but can be adapted to a range of endoscopic applications. This confirms the potential of learning-based monocular 3D reconstruction as a general framework for anatomical surface modeling in minimally invasive diagnostics and surgical planning.

Finally, **a general conclusion** summarizes the major contributions of this thesis and gives some perspectives which can improve the potential of the described work.

Chapter 1

Medical context and scientific objectives

1.1 Medical Context ◂

1.1.1 Endoscopy and Clinical Diagnosis ◂

Endoscopy is widely recognized as an essential diagnostic technique for identifying and monitoring lesions within hollow organs such as the esophagus, stomach, small intestine, bladder, and colon. This procedure allows for a direct visual (2D) investigation of the internal mucosa surfaces, and stands as a unique clinical method that captures the natural hues and detailed textures of epithelial tissues. Images are obtained using traditional Charge Coupled Device (CCD) cameras and provide crucial data for detecting and diagnosing lesions (e.g., cancers or inflammations) through textures, structures, and color information typically employed by the human brain for scene analysis.

Endoscopic procedures are generally non-invasive or minimally invasive since they avoid large incisions. Endoscopes are inserted into the body through natural openings or small cuts. These tools allow direct observation of both internal and external surfaces of various organs, including the lungs, liver, bladder, colon, esophagus, stomach, and joints.

The main objective of an endoscope is to capture images or video sequences of internal organ tissues. However, many endoscopes are also equipped with an operating channel, allowing to insert tools to perform biopsies or minor surgical interventions, such as gallbladder removal. Figure 1.1a outlines the general principle behind upper endoscopy (esophagogastroduodenoscopy, or EGD).

As illustrated in Figure 1.1, endoscopes are indispensable tools for exploring various anatomical regions. These include the gastrointestinal tract (esophagus, stomach, duodenum, colon, rectum, and anus); the respiratory tract, comprising the nasal cavity, paranasal sinuses, pharynx, trachea, bronchi, bronchioles, and lungs; the auditory system (outer, middle, and inner ear); the urinary tract (kidneys, ureters, bladder, and urethra); the female reproductive system (cervix, uterus, and fallopian tubes); the joints, such as the knees and elbows; and internal cavities such as the abdominal and pleural cavities (see Cappell [2008]).

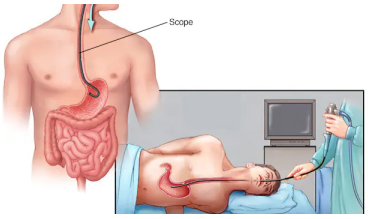
In some cases, endoscopic imaging can be combined with other modalities, such as ultrasound. For instance, an ultrasound probe may be mounted on the endoscope to capture supplementary information about the internal walls of the esophagus or stomach. This endoscopic ultrasound technique can also visualize hard-to-reach organs such as the pancreas.

A unique form of endoscopic investigation exists for the small intestine: *capsule video endoscopy*. This technique involves ingesting a small capsule that contains a miniaturized, wireless CCD camera, which records 5 to 10 images per second. These images are transmitted wirelessly to a

receiver worn by the patient or stored locally in the capsule for later retrieval. Capsule endoscopy is particularly useful in cases of chronic gastrointestinal bleeding when conventional endoscopies (gastroscopy and colonoscopy) have failed to identify the source.

Most standard endoscopes are composed of several key components. First, the insertion tube, which may be rigid or flexible, guides the distal tip to the area of interest. While rigid tubes are typically used in minimally invasive surgeries (e.g., abdominal or joint procedures), flexible tubes are preferred for examining hollow organs due to their enhanced maneuverability. Second, a lighting system, which usually employs a white-light source, is in charge of providing illumination through optical fibers from the proximal to the distal end, generating a diffuse light cone. In some specialized procedures, alternative illumination sources may be used. Third, the imaging system consists of optics and a CCD sensor mounted directly on the distal tip—a design known as *chip-on-the-tip*. This technology eliminates the honeycomb artifact seen in older fiber-optic systems, providing sharper, more contrast-rich images, especially when motion is moderate. The optics are designed with short focal lengths and wide apertures to maximize resolution, though this limits the field of view and can introduce barrel distortion. Finally, with chip-on-the-tip systems, electrical signals carrying image data are transmitted directly from the distal tip to the processing unit.

The dimensions of the endoscope—such as tube length and diameter—as well as its name, are determined by the specific organ or system being examined. For instance, arthroscopes, broncho-



(a) Upper gastrointestinal (GI) endoscopy

(b) Rigid cystoscope

(c) Flexible cystoscope

FIGURE 1.1 – Procedures and instruments in endoscopy. (a) Upper GI endoscopy is a procedure used to diagnose and, in some cases, treat lesions affecting the upper digestive tract, which includes the esophagus, stomach, and duodenum (the first part of the small intestine). A flexible endoscope is inserted through the mouth (serving as the natural orifice) and navigated along the upper digestive tract. In modern digital endoscopy systems, a CCD sensor is mounted at the distal tip (i.e., the portion inserted into the body), while the camera system is connected to the proximal end (only the eyecup, indicated by the "endoscope" label, is visible in the image). Biopsies can be performed through the instrument channel, as the passage from the mouth to the duodenum is typically unobstructed. (b)–(c) In urology, the endoscope is inserted via the external urethral orifice to inspect the bladder wall and urethra. This procedure may be conducted using a rigid endoscope (b), typically when surgical intervention is necessary, or with a flexible fiberscope (c), which offers greater maneuverability and improved patient comfort during routine examinations. The illustrations were adapted from (Phan [2020])

.

scopes, and nephroscopes are specialized for joints, the bronchial tree, and renal cavities, respectively. In contrast, gastroscopes and colonoscopes target the upper and lower digestive tracts, while cystoscopes and ureteroscopes are used for the lower and upper urinary tracts. Laparoscopes are a particular type of endoscope employed in minimally invasive surgical procedures. Consequently,e terminology of the examination reflects the organ or tract under observation, leading to terms such as gastroscopy, cystoscopy, or ureteroscopy.

Just as the external body is covered by skin, the internal surfaces of hollow organs—such as the esophagus, stomach, small intestine, colon, and bladder—are lined with epithelial tissue. The primary objective of endoscopic imaging is to visualize both healthy epithelial textures and pathological structures or signals indicative of lesions. Most standard endoscopic procedures use white light as the illumination source. This imaging modality enables a natural rendering of epithelial colors and textures, which assists endoscopists in identifying organ anatomy, recognizing key anatomical landmarks for navigation, and detecting certain types of lesions. Despite its widespread use, white light may not always reveal all types of lesions or allow for early-stage diagnosis. Therefore, many endoscopic systems provide the ability to alternate between different illumination sources: white light for realistic tissue visualization and initial diagnosis, and an alternative light source that enhances lesion visibility—especially at early stages—though often with reduced fidelity in natural appearance.

As illustrated in Figure 1.2, gastroscopy involves capturing close-up images of the gastric epithelium, which requires specific preparation to ensure optimal visualization. The patient must fast, avoiding food, drinks, and tobacco for at least six hours before the procedure, to clear the stomach.

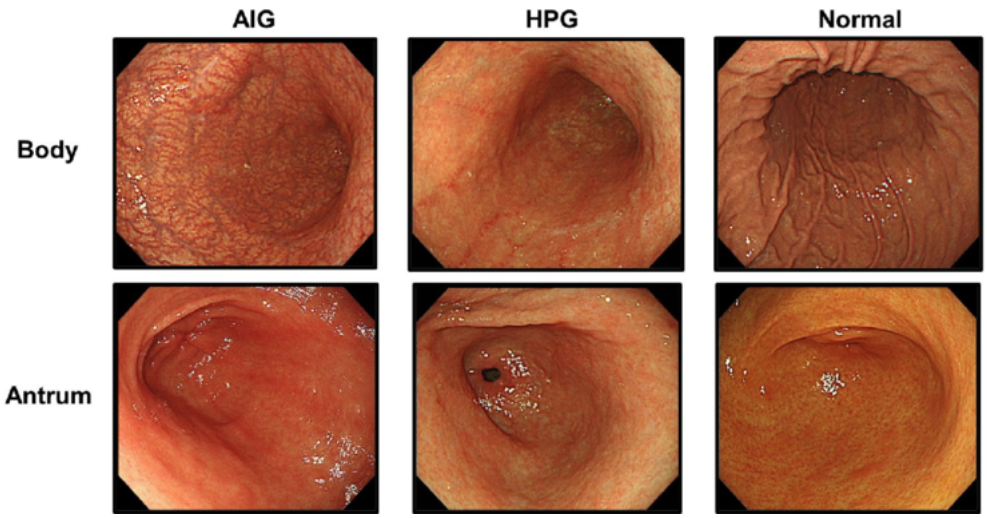


FIGURE 1.2 – Representative endoscopic images of the gastric body and antrum under three different conditions: Autoimmune Gastritis (AIG), Helicobacter pylori Gastritis (HPG), and normal gastric mucosa. Each column corresponds to a specific condition, while rows depict different anatomical regions of the stomach. Notable visual differences can be observed in mucosal texture and color, reflecting pathological changes associated with each condition.

During the examination, air is insufflated into the stomach to distend it and enhance visibility. However, despite the insufflation, the stomach walls remain flexible and subject to deformation, particularly due to the movements of the endoscope.

Two primary regions are typically monitored: the pyloric antrum, located in the lower portion of the stomach, and the gastric body near the cardia, close to the esophageal junction. These areas are clinically relevant, as they are prone to chronic inflammation that may lead to malignancy. The endoscopic image set previously shown includes representative white light (WL) images from these regions, comparing normal mucosa with pathological conditions such as Autoimmune Gastritis (AIG) and Helicobacter pylori Gastritis (HPG). White light imaging offers a natural visualization but tends to detect inflammation only at more advanced stages.

Narrow Band Imaging (NBI), which uses a green-blue light source, can be employed to enhance early lesion detection. Though less natural in appearance, NBI increases contrast in the epithelial texture, improving the detection of intestinal metaplasia and dysplastic lesions, especially in the antral region.

Similarly, in urological procedures such as cystoscopy, a comparable approach is used: the bladder is filled with isotonic saline to distend the organ and stabilize its walls, facilitating clearer imaging. Cystoscopic sequences are usually acquired under white light, although fluorescence imaging (FL) is also used, particularly for early detection of Carcinoma In Situ (CIS), which is typically invisible in standard white light. The FL modality penetrates deeper epithelial layers, enabling earlier identification of such hidden lesions, albeit at the cost of natural texture appearance.

Across these endoscopic modalities — either gastroscopy or cystoscopy— the common principle lies in the proximity of the camera to the tissue, enabling high-resolution image acquisition that is critical for early diagnosis.

Colonoscopy, like other endoscopic procedures, relies on the acquisition of high-resolution images in close proximity to the mucosal surface. To ensure optimal visualization, patients must undergo bowel preparation, typically consisting of a low-residue diet followed by laxative administration to clear the intestinal lumen. During examination, air or carbon dioxide is insufflated into the colon to distend its walls, facilitating navigation and improving mucosal exposure. However, the colon walls remain dynamic and deformable due to the peristaltic motion and the physical interaction with the endoscope.

Throughout the procedure, various anatomical regions—including the rectum, sigmoid, descending, transverse, and ascending colon—are examined for signs of pathology such as inflammation, polyps, or neoplastic lesions. Chronic inflammatory diseases such as ulcerative colitis and Crohn's disease are well-known risk factors for the development of dysplasia and colorectal carcinoma, thereby necessitating early and precise detection.

As illustrated in Figure 1.3, a representative sequence of images shows the evolution from lesion detection to endoscopic resection using enhanced imaging techniques.

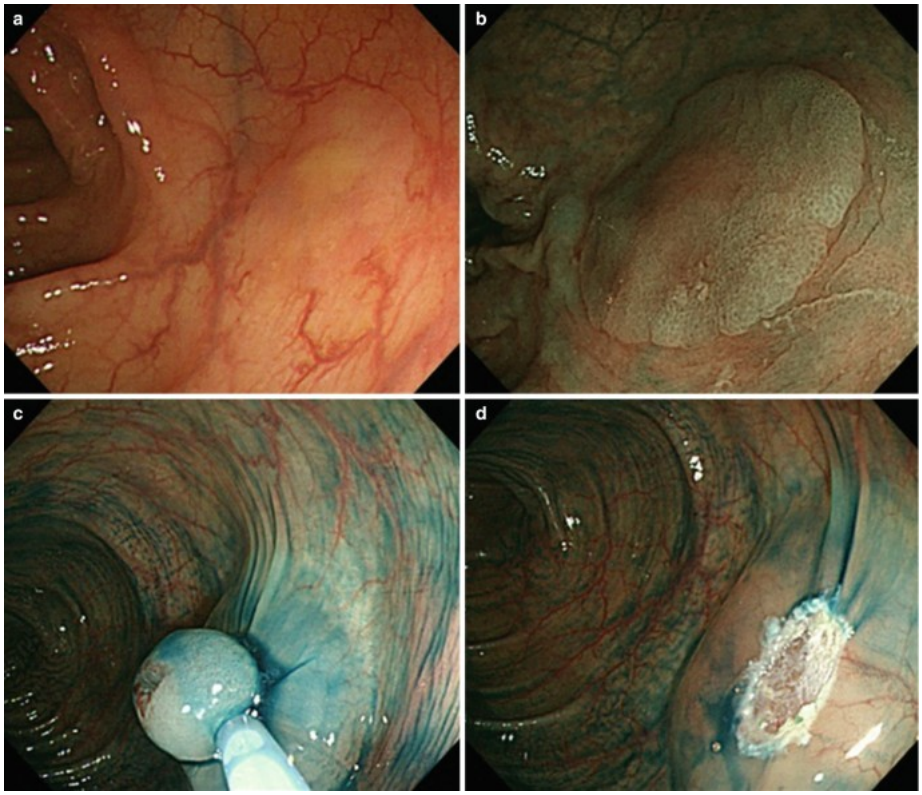


FIGURE 1.3 – Colonoscopic images showing the progression from lesion detection to resection using enhanced imaging techniques. (a) Initial view of the lesion under **white light (WL)**, displaying subtle mucosal irregularity. (b) The lesion under **Narrow Band Imaging (NBI)** reveals more distinct vascular and surface patterns. (c) Application of contrast dye during **chromoendoscopy** combined with NBI, highlighting the lesion margins before resection. (d) Post-resection view with visible submucosa and resection site. This sequence illustrates the clinical benefit of combining NBI and chromoendoscopy to improve visualization, delineation, and endoscopic treatment of colonic lesions.

These advanced imaging modalities, particularly NBI and dye-based chromoendoscopy, are increasingly integrated into routine colonoscopic practice, especially for the surveillance of high-risk patients. By enhancing the visualization of surface texture and vascular patterns, they significantly improve the detection, classification, and management of precancerous and early neoplastic lesions.

In conclusion, colonoscopy remains the reference standard for colorectal cancer screening and diagnosis. The integration of optical enhancement techniques, as illustrated in the figure, reinforces its efficacy in early lesion identification, accurate characterization, and targeted therapeutic intervention.

1.1.2 The Need for Extended Surface Representations ◄

1.1.2.1 Limitations of endoscopic images in terms of interpretability ◄

Unlike imaging systems that capture objects in wide field of view images acquired from a "rather large" distance, the distal tip of endoscopes is in close proximity to the epithelial surfaces. This proximity between the CCD-sensor matrix and the tissues imposes a fundamental limitation: the surface parts scanned by the clinicians are seen through narrow field of view (FoV) images. However, regions of clinical interest often span across tens or even hundreds of frames in a video sequence. A single frame provides only a fragmentary glimpse of the entire area of interest.

This constraint introduces several challenges from a medical point of view:

1. **Partial visibility of lesions:** Pathologies such as multi-focal cancerous lesions in cystoscopy or inflammations in gastroscopy are only partially visualizedin narrow FoV images. Such partial visualization of the region of interest hinders a comprehensive assessment and makes an accurate diagnosis difficult.
2. **Loss of spatial context:** In organs like the bladder, urologists cannot simultaneously visualize both lesions and key anatomical landmarks (e.g., urethra, ureters or air bubbles). To mentally represent the bladder in 3D, endoscopists must manipulate the scope with repetitive back-and-forth or zigzag motions, alternating views between the lesion and anatomical landmarks. This process is labor-intensive and time-consuming.
3. **Insufficient interpretability of video data:** A recorded endoscopic video, without the real-time handling of the instrument, often lacks sufficient contextual information for a proper scene interpretation. Consequently, in specialties such as urology or gastroenterology, routine recording of procedures is uncommon. This results in a dual limitation: there is no post-examination visual record for interdisciplinary consultation, and no traceable documentation of the procedure itself.
4. **Lack of longitudinal comparability:** As endoscopic video sequences are acquired at different times (e.g., weeks or months apart), they lack sufficient spatial overlap or standardization to allow meaningful comparison. As a result, evaluating the progression or regression of lesions over time is not possible.
5. **Incomplete mucosal coverage:** For hollow organs such as the stomach or the bladder, ensuring a thorough examination requires scanning the entire internal surface. With the limited FoV provided by the endoscope, it is difficult to guarantee full coverage, increasing the risk of missing significant lesions.

These limitations collectively emphasize the importance and clinical value of extending the field of view (FoV) in endoscopic imaging. A broader visual coverage would enable more accurate

diagnosis, improve scene understanding, facilitate the documentation of endoscopic examinations and make a patient follow-up possible.

1.1.2.2 Image mosaicing: 2D approaches versus 3D methods. ◄

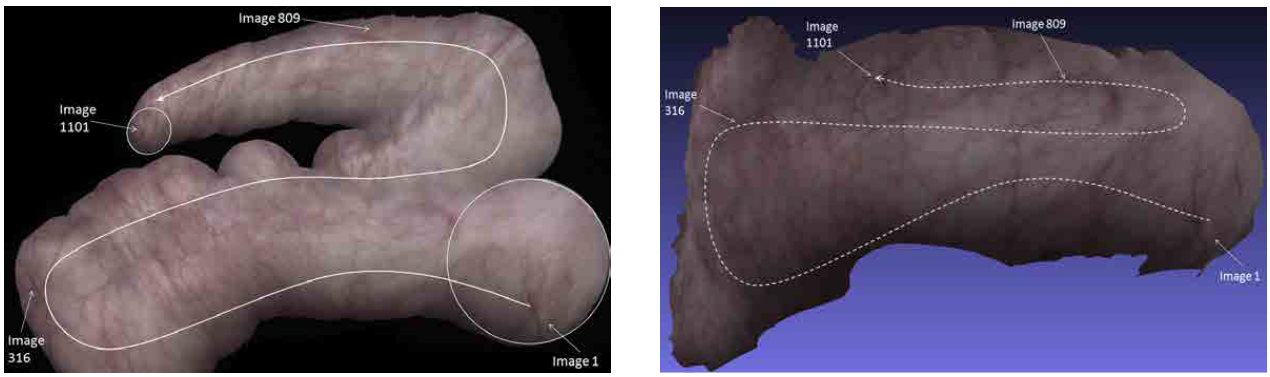
The field of view (FoV) in endoscopic imaging can be extended through the construction of panoramic views, which may be either two-dimensional (2D) or three-dimensional (3D). Regardless of the chosen technique, the resulting mosaics must offer a visually coherent rendering. In particular, they should avoid structural, textural, or color discontinuities that may arise from combining pixel information sourced from multiple frames into a unified coordinate system.

An important requirement of the mosaicing process is to preserve image resolution. When generating the mosaic from a sequence of images, resolution loss must be avoided: the final mosaic should ideally maintain the same resolution as the original input images across the entire extended field of view.

Over the past two decades, 2D mosaicing techniques have been actively explored in endoscopic applications. Notable contributions include work in urology for bladder wall mapping (Weibel et al. [2012b]; Behrens et al. [2009]) , gastroscopy for stomach inspection (Ali et al. [2016a]; Trinh et al. [2018]), and other endoscopic modalities as White-Light (WL) and Structured-Light (SL) (Seshamani et al. [2006]; Carroll and Seitz [2007]). The construction of a visually consistent 2D endoscopic mosaic typically involves the following key steps:

1. **Estimation of geometric correspondence:** The geometric relationship between successive image pairs is determined. This can be achieved either through global transformations, such as homographies that relate all pixels of one image to another (Weibel et al. [2012b]), or through dense vector fields, where each vector defines a correspondence between individual pixel pairs (Ali et al. [2016b]).
2. **Initial pixel mapping:** The estimated geometric transformations are then used to project and position the pixels from each image into a common mosaic coordinate system, forming an initial approximation of the panoramic reconstruction (Miranda-Luna et al. [2008]).
3. **Global optimization (bundle adjustment):** To refine the mosaic and reduce misalignment artifacts, global bundle adjustment techniques are applied. These methods adjust either the parameters of the geometric transformations or the pixel positions directly, with the goal of minimizing structural and textural discontinuities between overlapping images (Zenteno et al. [2022]).
4. **Correction of color inconsistencies:** Variations in illumination due to changes in endoscope orientation often lead to color discontinuities across the mosaic. These variations can be mitigated by either selecting pixel values that minimize visible differences or by applying weighted averaging across neighboring pixel colors to produce smoother transitions (Weibel et al.

[2012a]).



(a) (b)

FIGURE 1.4 – 2D and 3D mosaics examples generated from the same endoscopic image sequence. (a) 2D mosaic constructed from 1101 images (Ali [2016]), where the acquisition path is indicated by the white curve. Each image I_i is sequentially integrated in the mosaic, with visible resolution degradation due to varying endoscope viewpoints. Notably, the first (I_1) and last (I_{1101}) images exhibit substantial texture detail differences. Significant resolution loss and geometric distortion are observed due to strong viewpoint changes, except for the initial frame. (b) 3D mosaic reconstructed from the same sequence as that in (a), achieving full surface coverage without gaps, and preserving from texture discontinuities (Phan et al. [2020]).

1.1.3 Limitations of 2D Mosaics and Advantages of 3D Reconstruction ◂

In the specific case of *cystoscopy*, currently the most extensively explored domain for 2D endoscopic mapping, valuable insights have emerged from existing studies. In one of the first comprehensive works in this field (Miranda-Luna et al. [2008]), a sequential mosaicing strategy was proposed, where pixels from each newly registered image were iteratively added to a growing mosaic. However, despite the use of a robust and precise registration method based on the mutual information (Viola and Wells [1995]) acting as similarity measure, error accumulation along the video sequence resulted in visible inconsistencies. Texture misalignments and discontinuities can occur when extending the mosaic over a relatively modest area at images trajectory crossings in the 2D mosaic. Subsequent methods, including more sophisticated global correction algorithms (Weibel et al. [2012a]), were only capable of producing visually coherent mosaics for larger FoVs, at the expense of computation time.

More broadly, across various endoscopic applications, 2D mosaics do allow for an enlarged field of view, but they suffer from two critical limitations:

— **Projection distortion:** Since the human organ is three-dimensional, projecting its surface onto a 2D plane inherently introduces geometrical distortions. These become more pronounced as the mosaic extends farther from the reference image used as the reference plane for

projection. As a consequence, there is a noticeable loss of resolution and anatomical inaccuracy near the periphery of the mosaic, where the camera viewpoints are significantly different from that of the reference images defining the mosaic coordinate system. Figure 1.4(a) Illustrates this issue with an extended bladder mosaic showing severe resolution degradation and gaps in the epithelial surface.

— **Cognitive mismatch:** 2D mosaics do not align with the three-dimensional mental representation that clinicians naturally develop while navigating within the organ. This discrepancy reduces the interpretability and clinical utility of such representations.

In contrast, *3D mosaicing* offers a promising solution by addressing the core limitations of its 2D counterpart. When a 3D surface is reconstructed and image textures are accurately projected onto it using known camera poses, several advantages emerge:

— If the reconstructed surface geometry closely approximates the true shape of the organ, then image textures can be projected with relatively uniform resolution across the surface. This resolution remains largely independent of camera movement and viewpoint, and can even be enhanced using super-resolution techniques.

— Color discontinuities, often caused by varying lighting angles due to endoscope motion, can be better managed using geometric information. Specifically, the known local surface orientation can guide corrections, improving color consistency across frames.

— Texture gaps in 3D mosaics arise only in regions that were not scanned during acquisition, as opposed to 2D mosaics where such gaps often result from cumulative registration errors. This distinction is evident when comparing Figures 1.4 (a) and (b).

In summary, 3D mosaicing enables a substantial extension of the field of view while maintaining high-resolution texture quality and significantly improving visual coherence. Compared to 2D mosaics, 3D reconstructions provide better consistency in structure, texture, and color, offering a representation that is both more accurate and more intuitive for clinical interpretation.

1.2 3D Tissue Reconstruction in Endoscopy ◂

This section reviews the different 3D mosaicing methods proposed in the literature and their applications in different endoscopic procedures. While traditional approaches, often based on structure from motion, have demonstrated potential for 3D cartography of hollow organs, they present several limitations in clinical settings. In recent years, deep learning has emerged as a powerful alternative, offering new paradigms for 3D reconstruction that can overcome many of the challenges faced by classical pipelines. A brief overview of the structure-from-motion component in conventional methods is also provided to contextualize this transition.

1.2.1 3D Reconstruction and Mosaicing Principles ◂

3D reconstruction refers to the process of recovering the shape or structure of surfaces captured in images. This long-standing research domain remains highly relevant today due to its wide range of applications, which include computer-aided geometric design, computer graphics, animation, medical imaging, virtual reality, among others (Heinly et al. [2015]; Zhou et al. [2024]; Agarwal et al. [2009]). Depending on the context, the objective may vary from estimating precise object dimensions, as required in industrial metrology, to simply recovering surface geometry regardless of scale.

Traditionally, 3D reconstruction techniques have been classified into two main categories: active and passive vision methods. Active vision techniques rely on devices that project structured or controlled artificial light into the scene. The reflected signals are captured by a camera and used for surface reconstruction. Active vision devices used in endoscopy include Time-of-Flight (ToF) cameras (Lindner et al. [2010]; Penne et al. [2009]) and structured light systems (Ben-Hamadou et al. [2013]; Shevchenko et al. [2012]).

In contrast, passive vision methods reconstruct 3D surfaces using only images acquired from different viewpoints, without any external light projection. These methods include classical stereo vision, as well as more advanced techniques such as Simultaneous Localization and Mapping (SLAM, (Davison [2003])) , Shape-from-X methods (SfX, (Cryer et al. [1995]; Bregler et al. [2000]; Schönberger and Frahm [2016])), including Deformable Shape-from-Motion (DSfM, (Bartoli et al. [2012])), Structure-from-Motion (SfM, (Schönberger and Frahm [2016])), and Shape-from-Shading (SfS, (Wu et al. [2010])).

Both categories have led to accurate and robust 3D reconstructions across diverse domains such as industrial inspection (Graebling et al. [1995]; Dietrich [2016]) , robotics (Rossi et al. [2009]; Kuo et al. [2011]), and geological surveying (Zanchi et al. [2009]; James and Robson [2012]).

1.2.2 Active vision systems in 3D endoscopy ◂

One of the main advantages of active vision systems is that they do not require the detection and matching of homologous features, such as image primitives (e.g., corner or edge points) or textures to perform the 3D reconstruction. This is particularly beneficial in endoscopic scenes, where image primitives or textural cues may be absent or extremely challenging to extract. Instead, these systems rely on information derived from controlled light emitted by a projector.

3D reconstruction attempts in endoscopy with ToF systems. In the context of endoscopy, 3D reconstruction has been explored using Time-of-Flight (ToF) systems. ToF is an active sensing technique that measures the time taken for emitted light to travel to an object and return to a detector. This principle has been leveraged to develop a new generation of range-sensing devices based on conventional CMOS (Complementary Metal-Oxide-Semiconductor) or CCD (Charge-Coupled De-

vice) technology, which forms the basis of most ToF cameras (Maier-Hein et al. [2013]). These cameras typically comprise several components, including an illumination unit, optical elements, an image sensor, driver electronics, and a distance computation module (Hansard et al. [2013]). A schematic illustration of how a ToF camera reconstructs 3D points is shown in Fig. 1.5.

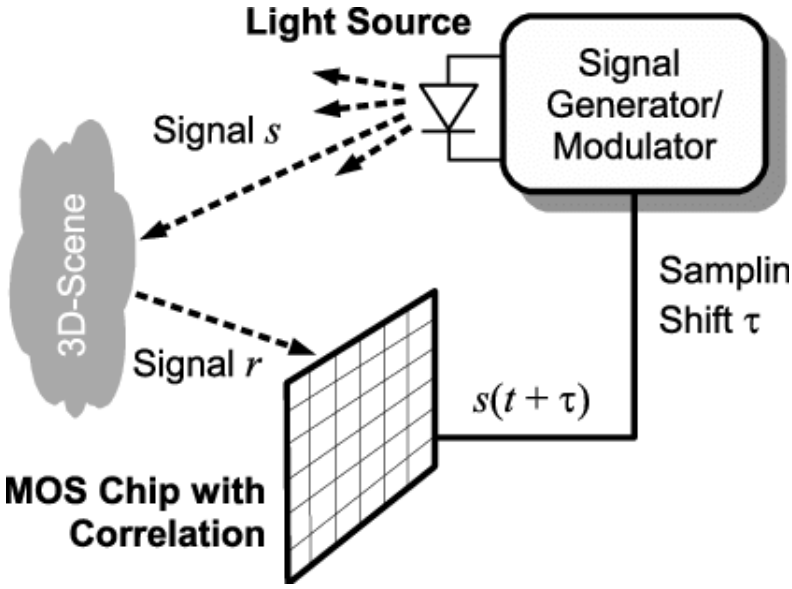


FIGURE 1.5 – The Photonic Mixer Device (PMD) delivers full-range distance data in real time by projecting modulated near-infrared (NIR) light onto the scene and measuring the phase shift between the emitted reference signal and the reflected light (Lindner et al. [2010]). The camera's illumination units emit intensity-modulated NIR light, synchronized with an internal reference signal. This light reflects off object surfaces and is captured by specialized smart pixels within the ToF camera. The accompanying illustration is adapted from (Lindner et al. [2010]).

Phase-based ToF systems face several inherent challenges that affect the accuracy of depth measurements. These include systematic calibration errors and issues arising from parameter interdependencies. One factor contributing to measurement inaccuracy is the physical implementation of light modulation, typically performed using LEDs. Instead of emitting ideal sinusoidal waves, the LEDs produce an approximation in the form of band-limited rectangular signals. This results in non-linear distortions in depth measurements, commonly referred to as "wiggling errors". Another common artifact is the presence of "flying pixels", which appear in regions with abrupt depth variations and lead to unreliable measurements (Foix et al. [2011]; Maier-Hein et al. [2013]).

Ongoing research seeks to mitigate these errors. For example, camera simulation techniques incorporating intensity-based calibration models, such as the one proposed by Lindner et al. (Lindner et al. [2010]), have been shown to reduce the impact of flying pixel artifacts.

A notable milestone in the application of ToF for endoscopy was the development of the first ToF-based endoscopic system by Penne et al. (Penne et al. [2009]). This system combined a commercially avalaible ToF camera (PMD[vision]3k-S by PMD Technologies, Siegen, Germany) with rigid endoscopic optics, capturing depth data at up to 25 frames per second with a resolution of 48 ×

64 pixels. The authors evaluated the system's accuracy on phantom models—specifically, an excised pig stomach—and reported an average depth measurement precision of 0.89 mm and a median precision of 0.71 mm, based on 100 static depth maps.

ToF cameras offer a significant benefit by delivering both depth and intensity data at high frame rates with relatively compact hardware (Maier-Hein et al. [2013]). However, their integration into endoscopic systems is fraught with limitations. The incorporation of ToF technology into conventional endoscopes necessitates substantial and costly hardware modifications. Additionally, the complex optical pathways in endoscopic imaging create non-uniform infrared illumination, which degrades the accuracy of range measurements and introduces significant noise (Haase and Maier [2018]).

Despite providing dense 3D point clouds, ToF systems still suffer from a narrow field of view and limited spatial resolution when compared to standard 2D endoscopic imaging. This mismatch necessitates the combined use of high-resolution 2D images and the comparatively less accurate 3D depth maps from ToF cameras—one of several reasons why ToF integration in endoscopy remains both expensive and technically complex.

3D reconstruction in endoscopy using structured light systems. Structured light systems project a known pattern (often grids or horizontal bars) onto the scene. These systems rely on the parallax between the optical axes of the camera and the projector. As illustrated in Fig. 1.6, triangulation techniques are then employed to reconstruct the position of 3D points.

The structured light system must first be calibrated to enable 3D measurements. This calibration is generally based on a mathematical model describing the projection of the pattern into 3D space, with all equations defined in the camera's coordinate system. A well-known method for calibrating camera/projector systems was proposed in (Ben-Hamadou et al. [2013]). Their procedure supports various point pattern configurations and does not depend on the number, color, or spatial distribution of the projected points. Furthermore, no external positioning device is needed since the projector geometry can be estimated directly within the camera's coordinate system using only unknown calibration board positions.

The principle behind this approach has also been applied to reconstruct small surface areas of the bladder in endoscopy, as demonstrated and validated on phantoms in (Shevchenko et al. [2012]). Building upon this idea, the CRAN laboratory introduced a method to expand the 3D field of view in endoscopic bladder imaging (Ben-Hamadou et al. [2016]). For each sensor pose, a point cloud is reconstructed in the camera's coordinate frame using the triangulation principle shown in Fig. 1.6. As the camera moves between two consecutive acquisitions i and $i+1$, a rigid 3D transformation (i.e., three translation parameters and rotations around three axes) $T_{3D}^{i,i+1}$ relates the two coordinate systems in which the 3D points coordinates are determined for a given viewpoint.

It is assumed that the geometric relationship between homologous pixels in consecutive images I_i and I_{i+1} can be modeled by a homography $H_{2D}^{i,i+1}$. The method in (Ben-Hamadou et al. [2016])

estimates the camera displacement between frames by iteratively optimizing the translation and rotation parameters of $T_{3D}^{i,i+1}$ so that the corresponding homography $H_{2D}^{i,i+1}$ maximizes a similarity metric (mutual information, see (Miranda-Luna et al. [2008])) determined with the grey-levels of the common regions of images I_i and I_{i+1}. The transformation that maximizes this measure is then used to align the two point clouds determined for camera poses i and $i+1$ into a common coordinate system.

By repeating this process over all images of a video-sequence, the field of view can be incrementally expanded by placing the 3D-points obtained for successive camera poses in the 3D coordinate system of the mosaic. The tests on realistic bladder phantoms reported in (Ben-Hamadou et al. [2016]) confirmed the feasibility of this structured light-based active vision system.

The main advantages of this approach lie in its high speed, accuracy, and robustness for reconstructing 3D points, even in the absence of image features. However, aligning point clouds into a

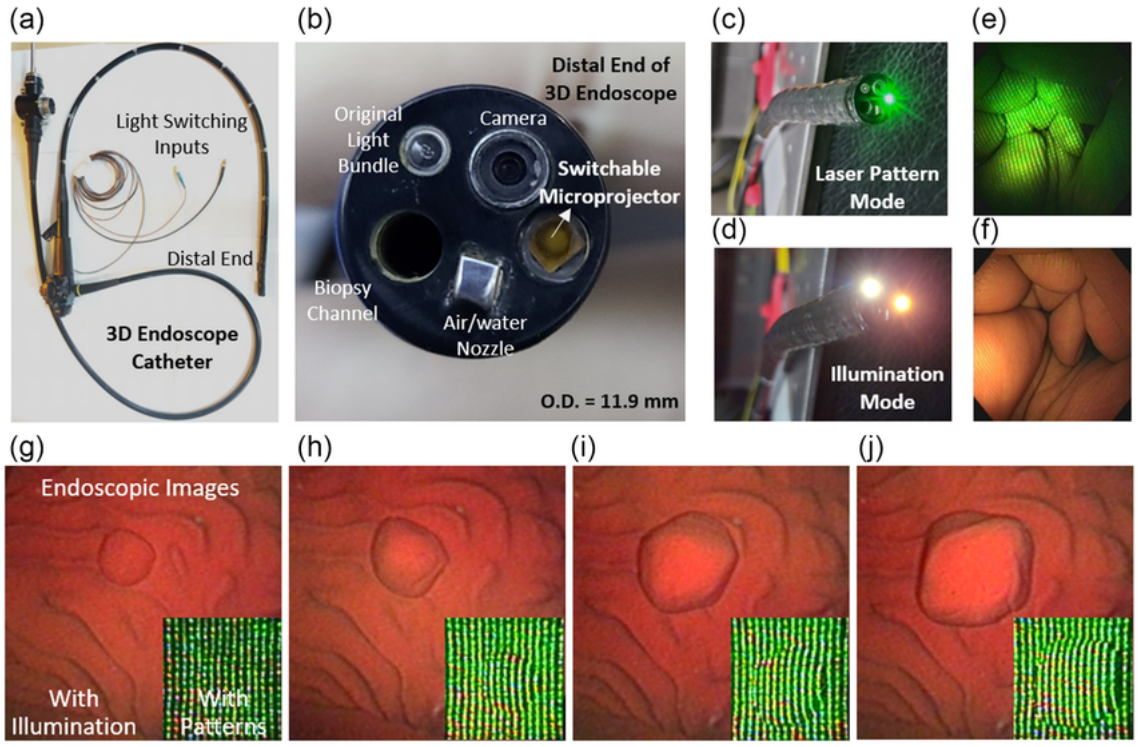


FIGURE 1.6 – Overview of a 3D endoscopic imaging system using structured light. (a) Complete 3D endoscope catheter with integrated light-switching inputs and distal end. (b) Close-up of the distal tip, showing key components including the camera, switchable microprojector, light bundle, biopsy channel, and air/water nozzle. (c) Activation of the laser pattern projection mode for structured light projection. (d) Standard illumination mode. (e) and (f) Endoscopic views with projected structured light pattern and with white light illumination, respectively. (g–j) Endoscopic images of tissue under different conditions, showing both illumination and projected pattern modes with zoomed insets highlighting the pattern details used for 3D reconstruction (Illustration adapted from (Yang et al. [2023]).

global coordinate system leads to cumulative errors along the trajectory, necessitating global surface correction. A more critical limitation of structured light systems is the significant hardware modifications required to integrate such systems in standard endoscopic systems. Compared to ToF systems, structured light solutions generally achieve a better 3D reconstruction accuracy, but more complex and expensive hardware upgrades are required, which significantly increases the cost of the endoscope.

1.2.3 Passive vision systems in 3D endoscopy ◂

Passive vision methods rely solely on the visual information contained in the acquired images. Thus, unlike active vision approaches, they offer the key advantage of keeping the hardware of existing endoscopic systems unchanged. All passive 3D reconstruction techniques are based on the principle of image content disparity, which arises from observing the scene from multiple viewpoints.

1.2.4 3D reconstruction based on stereoscopy. ◂

Stereoscopic reconstruction techniques rely on the geometric concept of parallax: the displacement between homologous points in two images taken from different viewpoints. Each pair of corresponding pixels defines a line of sight, and under ideal (numerical error-free) conditions, these lines intersect in the space at the actual 3D location of the scene point that is projected onto both image planes. Thus, each line represents the projection path of a 3D point into the respective image. The angle between these lines (or the baseline between the two camera optical centers) must be sufficiently wide to ensure accurate triangulation.

Classical stereoscopic 3D reconstruction algorithms follow a standard four-step procedure (Scharstein and Szeliski [2002]; Maier-Hein et al. [2013]; Daul et al. [2005]):

1. **Camera calibration:** Intrinsic and extrinsic camera parameters are estimated to enable the computation of projection rays from each image pixel.
2. **Feature detection:** Specific points of interest (or image primitives) are segmented in each image acquired from different viewpoints.
3. **Feature matching:** The correspondence of homologous points of interest is determined across the stereo images.
4. **Triangulation:** The 3D coordinates of the matched points are computed based on the intersection of the projection rays in a common scene coordinate system.

Stereoscopy has been widely used in laparoscopy-oriented 3D reconstruction methods (Stoyanov [2012]; Röhl et al. [2012]; Stoyanov et al. [2010]). In (Stoyanov et al. [2010]), the authors

introduced a semi-dense real-time reconstruction framework designed for minimally invasive surgery. Their approach begins with a sparse set of stereo correspondences and incrementally grows it into a semi-dense 3D point cloud using a best-first propagation algorithm (Lhuillier and Quan [2000]). The effectiveness of this method was validated on phantoms with known ground truth data, although no in vivo patient data was used for evaluating the method.

Stereoscopic imaging has long been recognized as a feasible approach for recovering 3D surfaces from in vivo laparoscopic images. It remains the most extensively tested technique in clinical settings, largely due to the integration of stereoscopic hardware in certain laparoscopes. These devices capture synchronized images from two slightly different viewpoints, either presenting them to the surgeon via dual monitors or stereoscopic glasses, or processing them for computational 3D reconstruction. Although depth perception in such systems is often handled by the human visual system, the image pairs can also be used for algorithmic depth recovery.

Despite these advantages, stereoscopy in endoscopy encounters significant obstacles. Factors such as illumination, reliance on varying viewpoints, specular reflections, intricate tissue shapes, and occlusions by surgical tools demand specialized stereo matching algorithms specifically designed for minimally invasive surgery. In addition, achieving successful reconstruction of parts of the 3D scene frequently results in restricted surface coverage.

It is also important to note that not all stereoscopic laparoscopes are suitable for computational 3D reconstruction. Some systems generate stereo images using beam-splitting optics, which lack a physical baseline between the two viewpoints—a critical component for triangulation. Additionally, all intrinsic parameters of the laparoscopic system (e.g., focal length, optical center, distortion parameters) must be calibrated prior to the procedure. Even with proper calibration, these parameters must remain constant throughout the operation. However, in practical laparoscopic procedures, the distance between the distal tip of the endoscope and the tissue frequently changes. This variability often requires focal adjustments to maintain sharp imaging, which in turn alters both the focal length and the distortion parameter values. Such changes negatively impact the accuracy of the reconstructed 3D points (Maier-Hein et al. [2013]).

1.2.5 3D reconstruction based on Shape from Shading (SfS). ◂

The concept of Shape-from-Shading (SfS) was introduced by Horn in the early 1970s (Horn [1975]), showing that the three-dimensional shape of a surface can be inferred from a single image. SfS techniques rely on models that relate pixel gray-level values to the corresponding surface normals. The authors in (Zhang et al. [1999]) proposed a taxonomy that groups the SfS approaches into four major categories of methods:

— The **energy minimization methods** reconstruct surfaces by optimizing an energy functional (Zhang and Cryer [1991]),

— The **propagation-based approaches** spread the surface information from known feature points across the image (Bruckstein and Netravali [1992]),

— The **local methods** use prior assumptions about the surface shape, such as spherical surfaces in (Lee and Rosenfeld [1985]),

— The **linearization techniques** solve a simplified version of the reflectance map using linear methods (Penedo [1988]).

Most SfS algorithms operate under the following assumptions: (i) a single light source is present, (ii) surfaces exhibit Lambertian reflectance (i.e., uniform reflectance in all directions), (iii) the surface albedo is constant or known, and (iv) the surface is smooth and continuously differentiable (Maier-Hein et al. [2013]).

SfS has been successfully applied in various endoscopic procedures, including bone surface reconstruction (Wang et al. [2010]), real-time 3D visualization from monocular laparoscopic video (Cheng and Alkaisi [2012]), and capsule endoscopy (Pahlavan et al. [2012]). The authors in (Yeung et al. [1999]) proposed an SfS method based on the detection of singular points (e.g., maxima, minima, saddle points of the image brightness function and used level set propagation to derive distance maps from each surface point to the light source. These maps were aligned using known homologous points and back-projected to recover the 3D surface.

In a similar direction, the authors in (Forster and Thiran [2000]) proposed a technique for reconstructing portions of the stomach surface from single endoscopic images. The method assumes a spherical projection model for the camera and uses calibration data to estimate and correct image distortion. To address the challenge of specular reflections commonly present in endoscopic scenes, the authors employed a dichromatic reflection model to separate the diffuse and specular components, thereby approximating the Lambertian reflectance assumption required for SfS algorithms. However, the study did not include any quantitative evaluation of the reconstruction accuracy.

A key benefit of SfS is that it can be applied without modifying standard endoscopic hardware. However, these methods are limited by their strong assumptions regarding the underlying lighting conditions, surface reflectance, and geometric regularity. Consequently, SfS typically enables the reconstruction of only small and smooth surface regions.

1.2.6 3D Reconstruction Based on Structure-from-Motion (SfM) ◂

Structure-from-Motion (SfM) is a powerful technique for reconstructing the 3D shape of surfaces from a sequence of 2D images taken from different viewpoints. The goal of SfM is to recover the shape of a static (non-deformable) scene while simultaneously estimating the camera motion. This requires a sufficient number of accurate point correspondences across images, typically represented as point tracks. These point tracks are then used to infer both the trajectory of the camera and the 3D coordinates of surface points (James and Robson [2012]; Schönberger and Frahm [2016]).

Soper et al. (Soper et al. [2012b]) proposed a SfM approach to reconstruct the inner surface of the bladder wall using video data captured with a highly flexible ultrathin endoscope guided by a robotic system. The setup ensured ideal acquisition conditions: the optical axis was kept perpendicular to the surface, and images were acquired along a spiral trajectory, maximizing surface coverage and overlap. Although this configuration demonstrated promising results on a pig bladder phantom, it does not reflect the constraints of real clinical environments. Standard cystoscopes—whether rigid or flexible—cannot provide such precisely controlled acquisition conditions.

Building on this direction, the authors in (Lurie et al. [2017a]) proposed a dense SfM approach for bladder wall reconstruction using white light cystoscopy images obtained from conventional clinical systems. Their method relies on the assumption that a sufficient number of reliable correspondences can be detected across most images using the SIFT (Scale Invariant Feature Transform) algorithm (Lowe [2004]). However, this assumption is often violated in endoscopic imagery, where large regionsmay exhibit poorly textured regions, due to previous surgical procedures or after the patient received radiotherapy.

While both studies (Soper et al. [2012b]; Lurie et al. [2017a]) confirm the feasibility of applying SfM to endoscopic videos, their reliance on ideal acquisition setups or a general lack of textural features severely limits the clinical applicability of such approaches. In particular, gastroscopic images frequently suffer from a lack of texture and significant illumination variations between frames (Phan et al. [2019]). These conditions impair the effectiveness of feature detection and matching techniques, making it difficult for current SfM algorithms to robustly reconstruct 3D geometry from such data. Consequently, SfM methods cannot be applied to challenging endoscopic scenarios such as those found in gastroscopy.

To address some of these limitations, the study in (Phan et al. [2020]) proposed an optical flow-based SfM framework specifically designed for epithelial surface reconstruction in endoscopic imaging (see figure 1.4(b)). By leveraging dense motion estimation instead of sparse keypoint matching, their method enhances robustness in low-texture regions and under complex illumination changes—two common issues in real clinical data. This approach demonstrates that integrating optical flow with SfM can improve the reliability of 3D reconstructions in minimally textured and dynamically illuminated environments typical of endoscopic procedures. However, this robust and accurate surface reconstruction involves high computation time impeding the use of the method during the examination.

1.2.7 Deep Learning in 3D Reconstruction ◂

Recent advances in deep learning architectures have significantly transformed the landscape of 3D reconstruction in endoscopy, offering alternatives to classical geometry-based methods that struggle with the complex visual conditions found in clinical imagery. Estimating depth from indi-

vidual images is a fundamental step in 3D reconstruction and mosaicing algorithms. Precise depth estimation plays a critical role in enhancing the creation of reliable, accurate, and real-time SfM or SLAM algorithms. Depth estimation based on deep learning approaches can be broadly categorized into three main families: (i) **supervised learning methods**, which rely on depth or pose ground truth (often obtained from synthetic datasets or SfM pipelines) to train depth estimation networks; (ii) **self-supervised learning methods**, which circumvent the need for ground truth by exploiting geometric and photometric consistency across consecutive video frames; and (iii) **end-to-end implicit scene representation methods**, such as Neural Radiance Fields (NeRF) and Gaussian Splatting, which directly learn volumetric or point-based representations from posed images.

1.2.8 Supervised Depth Estimation ◂

Early deep learning-based approaches in the field of 3D reconstruction of endoscopic data often relied on supervision signals provided by SfM generated data or synthetic datasets.

In (Widya et al. [2019]), the authors used classical SfM to reconstruct the inner stomach surface using chromo-endoscopy (CE) images (in the CE-modality SIFT of SURF algorithms provide enough point correspondences for the SfM approach). Given the limited clinical adoption of CE, the same authors proposed a method in (Widya et al. [2021]) that translates white-light endoscopic images to virtual CE images using a CycleGAN (Zhu et al. [2017]). These synthetically enhanced images were then used for SfM-based reconstruction, demonstrating improved textural information but retaining the limitations of classical SfM, such as non real-time performance.

The authors in (Ma et al. [2019, 2021]) introduced RNNSLAM, a recurrent neural network-guided SLAM pipeline. Their method predicts depth and pose maps from colonoscopy frames and uses these prediction in a bundle adjustment steps to reconstruct the inner surface 3D colon parts. While this approach supports real-time inference, its accuracy and scalability are limited due to the dependency on SfM-based ground truth.The "correctness" of this ground truth allowed only to reconstruct small colon parts, the diameter of the colon chunks becoming quickly too large and unrealistic for longer tubular surfaces.

The work presented in (Mahmood and Durr [2018]) introduced a fully supervised monocular depth estimation approach based on a hybrid CNN–CRF architecture. In this context, a Conditional Random Field (CRF) is a probabilistic graphical model used to enforce spatial consistency in the predicted depth maps by modeling contextual dependencies between neighboring pixels. The network was trained using synthetic colonoscopy data and validated on a porcine colon dataset. The authors employed adversarial domain adaptation techniques to address the domain shift between synthetic and real-world images. Ground truth depth was obtained by registering endoscopic video frames with corresponding CT scans. Despite these efforts, the method's reliance on synthetic data for training limits its generalization, as such datasets often fail to capture the complex

visual phenomena—such as specular highlights, tissue deformation, and abrupt illumination variations—commonly encountered in real clinical endoscopic procedures.

These limitations in ground truth acquisition and the gap between synthetic and real data have spurred interest in self-supervised approaches.

1.2.9 Self-Supervised Depth Estimation ◂

Self-supervised methods forgo the need of explicit ground truth by leveraging geometric and photometric consistency relations across frames. Many existing self-supervised techniques, originally developed for urban or indoor scenes (Godard et al. [2019]; Zhou et al. [2017]; Diamantas et al. [2010]), assume brightness constancy (i.e., constant illumination conditions from one image to another). However, this assumption is often violated in endoscopic videos due to strong and complex lighting variations between different acquisition. These illumination changes are both related to viewpoint changes and to the change of distance between the endoscope's distal tip and the organ tissue. Moreover, within an image, the illumination changes may be local, i.e., different from one image part to another according to local surface orientations. "Naive" photometric loss functions are not suitable in these settings, as they do not take these complex (and local) illumination changes into account.

In (Li et al. [2021]), an LSTM-enhanced pose estimation network (EgoMotion) was introduced for performing depth prediction in laparoscopic scenes from the SCARED dataset (Allan et al. [2021]). While the approach is temporally-aware, it is still limited by the adopted brightness constancy assumption, particularly when the images include overexposed regions or specular reflections.

The authors in (Liu et al. [2019]) proposed a geometry-aware loss based on sparse motion flow and depth consistency to deal with the challenging acquisition conditions in endoscopy. Although the approach is robust to lighting variations, it requires an initial sparse reconstruction, which increases the computational cost and complexity.

The authors in (Ozyoruk et al. [2021]) introduced an affine transformation-based photometric model coupled with an attention-based pose predictor, which improves robustness to lighting variations. However, assuming a global affine model to approximate complex local illumination changes remains overly simplistic for highly dynamic endoscopic scenes.

Finally, the contribution in (Shao et al. [2022]) describes a more refined approach based on appearance flow estimation, enabling the model to account for significant local lighting variations. By explicitly predicting optical flow between frames, the method localizes homologous pixels. The knowledge of homologous point pairs predicted by the optical flow is used as a first step by a training phase to correct local illumination variations between images. However, the accuracy and robustness of the optical flow prediction method being themselves affected by complex illumination

changes, the depth prediction remains inaccurate, notably in the case of large camera motion as arising in endoscopy. Moreover, this method is computationally expensive due to its multi-stage training process. It first requires the pretraining of an optical flow estimation network, followed by a fine-tuning step involving the complete pipeline. This added complexity not only increases training time but also necessitates additional memory and computational resources.

1.2.10 End-to-End Implicit 3D Representations ◂

Recent advancements in neural rendering have introduced fully end-to-end approaches capable of learning continuous 3D scene representations from input images acquired from known (i.e., determined) camera poses, thereby enabling accurate 3D reconstruction. Notably, Neural Radiance Fields (NeRF) have been adapted to medical imaging applications, allowing for photorealistic and geometrically consistent reconstructions. For instance, the authors in (Wang et al. [2022b]) proposed EndoNeRF, a framework for stereo 3D reconstruction of deformable tissues in robotic surgery, demonstrating its effectiveness in handling non-rigid deformations and occlusions. Similarly, the contribution in (Qin et al. [2024]) decribes the Endoscope-NeRF model which reconstructs implicit radiance fields of endoscopic scenes under non-fixed light sources, enhancing the realism of virtual surgical simulations.

Liu et al. introduced EndoGaussian, a real-time endoscopic scene reconstruction framework built on 3D Gaussian Splatting to address the challenges of dynamic scenes and real-time rendering. Their method (Liu et al. [2024]) achieves significant improvements in rendering speed and quality, making it suitable for intraoperative applications. Additionally, Bonilla et al. presented Gaussian Pancakes, which integrates geometrically-regularized 3D Gaussian Splatting with a Recurrent Neural Network-based SLAM system, resulting in smoother 3D reconstructions with detailed textures (Bonilla et al. [2024]).

These methods implicitly encode shape and appearance in neural fields or Gaussian primitives and require accurate pose estimation. Their application to endoscopy is still emerging, but preliminary research shows that they can produce smooth, high-fidelity reconstructions, even in low-textured regions and under varying illumination. However, deploying such methods in endoscopy faces several challenges: (i) obtaining accurate camera poses remains difficult without additional sensors; (ii) these methods require dense image sampling from multiple views; (iii) their high computational demands currently hinder clinical deployment. Nonetheless, the adaptability of implicit representations makes them promising for future real-time and data-efficient endoscopic 3D reconstruction systems.

1.3 Photometric Constancy as a Prerequisite for depth estimation and surface construction ◂

In the field of monocular endoscopic image data processing, the robustness of both classical and learning-based methods heavily depends on the assumption of photometric constancy across image sequences. This assumption is frequently violated in endoscopy, where uncontrolled lighting, proximity-induced exposure artifacts, and specular reflections introduce significant variability across frames. These effects degrade the image quality and directly impair reconstruction pipelines, particularly those based on Simultaneous Localization and Mapping (SLAM) or Structure-from-Motion (SfM), which rely on brightness constancy and geometric consistency throughout the reconstruction process.

Image enhancement methods for attenuating the effects of various sources leading to inconstant scene illumination have emerged as a critical pre-processing step To mitigate these issues. Effective photometric correction not only improves visual image quality, but also improves the efficiency of key treatment stages such as homologous data correspondence estimation, depth prediction, and pose refinement. Two main families of image enhancement approaches are considered in the literature: the one based on classical image processing techniques and the deep learning-based strategies.

1.3.1 Traditional Image Enhancement Methods ◂

In endoscopy, the first reason of illumination changes between images comes from the vignetting effect which is due to the fact that the light source irradiates more the image center than the image periphery (i.e., a surface with a constant albedo is brighter in the image center than in the periphery). This non-uniform illumination explains why brightness constancy is violated when a same scene part is observed for different viewpoints. The vignetting effect can be corrected by approximating the illumination profile by a polynomial during a calibration stage and by using it to correct the gray-levels (Doutre and Nasiopoulos [2009]). Another approach is to convolve the images with a low pass filter (the vignetting mainly includes low frequency components) and to subtract this background image from the acquired images (Miranda-Luna et al. [2004]). Even if these methods are efficient they do not compensate for non-brightness constancy due to local and changing surface orientations and various acquisition distances.

Other methods try to attenuate all effects leading to non-brightness constancy (i.e., whatever their origin) through global or local transformations of pixel intensity distributions. Among the techniques benchmarked in this work , two belong to the histogram-based class, namely the Range-Limited Bi-Histogram Equalization (RLBHE) and the Flattest Histogram Specification with Accurate Brightness Preservation (FHSABP) methods, while the remaining two rely on illumination modeling. The latter are the LIME (Low-Light Image Enhancement) and DUAL (a dual-illumination

estimation) methods. The RLBHE method enhances contrast by dividing the image histogram in sub-regions and by equalizing these sub-regions (Zuo et al. [2013]). This approach attenuates illumination changes while preserving the image brightness. FHSABP focuses on maintaining brightness while flattening the histogram to avoid over-enhancement (Wang et al. [2008]). In contrast, LIME estimates an illumination map to suppress dark regions while avoiding noise amplification (Guo et al. [2017a]). DUAL applies a two-path estimation of scene illumination and reflectance to correct under- and overexposed regions independently (Zhang et al. [2019a]).

Although these methods offer a lightweight and interpretable solution, they suffer from limitations when dealing with the simultaneous presence of over- and underexposures in a same image — a common occurrence in endoscopic imagery.

1.3.2 Deep Learning-Based Image Enhancement ◂

Deep learning-based methods provide a more adaptive and context-aware alternative in comparison to classical enhancement techniques. Unlike global pixel remapping approaches, these methods learn data-driven representations that enable the restoration of spatial detail while maintaining temporal consistency across frames. Zhang et al. (Zhang et al. [2021]) introduced a recurrent neural network that predicts adaptive gamma correction maps using synthetically generated training pairs. The temporal modeling capacity of the RNN enables consistent photometric enhancement throughout video sequences, reducing flicker and correcting localized illumination disturbances.

In a complementary direction, the authors in (Daher et al. [2023]) proposed **Endo-STTN**, a Spatial-Temporal Transformer Network designed to address the inpainting of specular reflections in endoscopic video. This architecture leverages a transformer-based attention mechanism to establish temporal correspondences across frames and to ensure structural coherence in the inpainted regions. The model was shown to improve both visual continuity and surface quality in downstream reconstruction tasks, particularly in sequences severely affected by reflective artifacts. Together, these contributions demonstrate the growing potential of temporally-aware deep learning models for enhancing photometric stability and structural integrity in endoscopic imagery.

However, a fundamental limitation in training supervised deep learning-based enhancement models lies in the absence of paired datasets, i.e., image sequences exhibiting both photometric degradation and corresponding high-quality references. In clinical settings, acquiring such pairs is impractical due to the dynamic nature of tissue appearance, variable lighting conditions, and the inability to reproduce identical views under different illumination settings. The **Endo4IE** dataset is introduced in Chapter 2 of this thesis to address this issue. This benchmark provides synthetic but realistic pairs of underexposed, overexposed, and reference-quality colonoscopic images, generated using an image-to-image translation pipeline applied to frames extracted from clinically acquired videos. The Endo4IE dataset enables supervised training of photometric correction models

under controlled and reproducible conditions. Based on this dataset, two learning-based methods are proposed and evaluated in Chapter 2. The first, **Endo-LMSPEC**, employs a multiscale Laplacian pyramid decomposition coupled with SSIM-based regularization terms to enhance global and local structures in underexposed and overexposed regions. The second, **Endo-ViT**, leverages a Vision Transformer architecture combined with histogram-aware and structure-preserving losses to achieve high-fidelity correction while maintaining real-time inference capabilities.

1.3.3 Impact of Lighting Enhancement on Colonoscopy 3D Reconstruction ◂

Photometric enhancement is a critical enabler for reliable monocular 3D reconstruction. Innovations such as GAN-driven illumination transfer (Cheng et al. [2021], Wang et al. [2022a]) and CNN-assisted monocular enhancement for disparity approximation (Luo et al. [2019]) underscore the importance of image pre-processing in modern reconstruction pipelines, where photometric inconsistencies can propagate and degrade downstream geometric fidelity. Zhao et al. (Zhao et al. [2021]) provided empirical validation for this paradigm, evaluating the integration of a gamma-based recurrent enhancement module into a unified RNN-SLAM framework (Ma et al. [2019]). This study assessed the influence of photometric correction on camera tracking and depth estimation during colonoscopic reconstruction, using Absolute Pose Error (APE) and Root Mean Squared Error (RMSE) against COLMAP-derived reference trajectories. In the visually unstable and texture-deficient domain of endoscopy, the preservation of photometric consistency across frames proves essential for robust and repeatable 3D reconstruction. In Chapter 2, a systematic investigation is conducted to quantify the impact of three distinct deep learning-based enhancement methods—Endo-LMSPEC, Endo-STTN, and a recurrent gamma correction network—on the accuracy of camera trajectories and the quality of 3D surface reconstructions in colonoscopy. Each model is incorporated as a pre-processing module into the RNN-SLAM pipeline, and their effects are analyzed through both quantitative trajectory metrics and qualitative depth map evaluations. Implementing these correction strategies early in the reconstruction process enhances geometric accuracy, diminishes drift, and leads to more dependable anatomical mapping—vital for clinical applications in diagnosis, treatment planning, and surgical guidance.

1.4 Global Discussion about 3D Endoscopy ◂

The primary objective of this thesis is to develop a 3D mosaicing framework tailored to endoscopic imaging, under the assumption that the surfaces as almost rigid in the video sequence used for the cartography. Achieving robust and accurate extended field-of-view (FoV) reconstructions critically depends on the choice of a suitable 3D reconstruction paradigm.

Active vision techniques, which rely on projecting structured or modulated light into the scene, have demonstrated strong performance in industrial settings, largely due to their resilience to challenging illumination and absence of textures. These systems, such as those based on laser triangulation or time-of-flight (ToF), benefit from well-controlled calibration and are capable of accurate geometric reconstruction even under poor visual conditions. However, their integration into clinical endoscopy remains highly problematic. Not only do they require extensive and costly hardware modifications to be embedded within endoscopes, but they also typically produce only localized surface reconstructions per acquisition. Attempts to implement laser- or ToF-based endoscopic systems have revealed significant challenges in terms of stability, scalability, and practical deployment. Additionally, extending the reconstructed surface across sequences remains an unresolved issue for most active vision solutions. Given these constraints, this thesis deliberately opts for a **passive vision** approach.

Passive vision methods infer 3D information solely from 2D image sequences, eliminating the need for specialized hardware. Among these, **Structure-from-Motion (SfM)** and its real-time counterpart, **Simultaneous Localization and Mapping (SLAM)**, provide inherently scalable solutions for reconstructing extended organ surfaces. Their scalability lies in their ability to incrementally process long sequences and large anatomical areas without requiring changes to the underlying algorithmic structure. By leveraging point correspondences across multiple frames, these methods jointly estimate both camera motion and 3D surface geometry. When the observed scenes contain sufficient texture or structural detail, they can produce accurate and dense reconstructions, even under uncalibrated imaging conditions. Such approaches have demonstrated promising results in endoscopic settings, particularly in urological applications (Lurie et al. [2017b]; Soper et al. [2012a]; Phan et al. [2020]).

In contrast, early passive methods such as **Shape-from-Shading (SfS)** and **stereoscopy** are more constrained in their applicability. SfS attempts to reconstruct 3D surfaces from a single image but relies on strict assumptions: Lambertian surface reflectance, a single fixed light source, and smoothly varying illumination—conditions that are rarely satisfied in the reflective and dynamically lit environment of endoscopy. Stereoscopy, on the other hand, requires precise calibration and a stable scene geometry across stereo views. It is typically limited to specialized dual-camera setups such as stereo-laparoscopes, which are not widely adopted and lack generalizability across different types of endoscopic procedures.

In recent years, **deep learning-based approaches** have emerged as a compelling alternative to classical geometric methods. These methods can be categorized into supervised, self-supervised, and end-to-end scene representation techniques. Supervised approaches, while powerful, require large sets of annotated data, obtained either by using synthetic data or by generating ground-truths using SfM. However, synthetic data are susceptible to domain shift when applied to real clinical data, and SfM approaches provide ground-truths affected by errors. Self-supervised learning alle-

viates the need for ground truth by relying on geometric consistency or photometric loss functions. However, these losses often assume brightness constancy or smooth motion, assumptions that are not always valid in the presence of specularities, organ deformations, and abrupt lighting changes.

Furthermore, a growing body of work highlights the importance of **image enhancement as a pre-processing step** for deep learning-based 3D reconstruction in endoscopy. Because endoscopic scenes frequently suffer from under- or overexposed regions, traditional SLAM and learning-based systems often fail to establish reliable correspondences. Zhang et al. (Zhang et al. [2021]) demonstrated that local, learning-based exposure correction significantly improves the quality of camera trajectory and depth estimation in RNN-SLAM pipelines. This allows the model to capture fine-grained illumination variations and maintain temporal consistency across frames. Their results highlight that pre-processing steps remain a vital component in ensuring the robustness and reliability of downstream 3D reconstruction tasks.

The most recent advances in deep learning, such as **Neural Radiance Fields (NeRF)** and **Gaussian Splatting**, offer new paradigms for volumetric and implicit scene reconstruction. These techniques can model complex non-rigid scenes and are capable of producing high-fidelity 3D models. However, they still depend on accurate camera poses and dense view sampling, and their high computational demands currently limit their clinical feasibility.

Taking into account all these considerations, i.e., (hardware limitations, the need for robustness to clinical variability, the crucial role of photometric quality, and the challenge of working without ground truth (no hyphen when used as a noun), the strategy adopted in this thesis centers on a self-supervised learning approach for depth estimation. Unlike classical geometric methods such as Structure-from-Motion (SfM) or fully supervised deep learning approaches that require extensive synthetic or annotated datasets, self-supervised models learn directly from raw image sequences by leveraging photometric consistency. This paradigm offers a flexible and scalable compromise, particularly suited for endoscopic scenarios where collecting ground-truth depth is often impractical due to anatomical variability, dynamic tissue deformation, and unstable lighting conditions.

The use of self-supervised learning for depth estimation enables training models directly on in vivo endoscopic videos by exploiting temporal continuity and enforcing geometric consistency across adjacent frames. While these methods still face challenges, such as robustness to large viewpoint changes and motion blur, their scalability, hardware independence, and ability to generalize across different anatomical scenes make them a promising foundation for building dense and extended 3D mosaics in hollow organs. Therefore, the methodological core of this thesis lies in the development and analysis of a self-supervised depth estimation-based 3D reconstruction pipeline tailored to the clinical constraints and variability of endoscopic imaging.

1.5 Thesis Objectives ◂

1.5.1 Scientific Objectives of the Thesis ◂

The central scientific objective of this thesis is to advance the 3D reconstruction of internal organ surfaces from monocular endoscopic video-sequences by integrating deep learning techniques into a passive vision framework. Traditional SfM pipelines, while effective in textured and rigid environments, struggle under the challenging visual conditions of endoscopy—characterized by specular reflections, dynamic lighting, limited texture, and tissue deformation. These limitations hinder the robustness and scalability of conventional reconstruction methods in clinical practice.

To overcome these challenges, this thesis proposes a hybrid approach that augments classical geometric reconstruction with deep learning-based modules. These modules address key limitations of passive vision pipelines by improving input image quality, predicting reliable depth information, and enabling more consistent pose and structure estimation.

The specific scientific objectives of the thesis are as follows:

— **The design of deep learning-based pre-processing techniques** focuses on locally and globally enhancing image quality by correcting illumination variations, improving contrast, and mitigating exposure-related artifacts. These enhancements aim to optimize the photometric conditions necessary for accurate depth prediction and, consequently, for reliable 3D surface reconstruction in endoscopic imagery.

— **The development of a self-supervised depth estimation network** tailored to monocular endoscopic sequences. The network is trained without ground truth depth supervision and is designed to handle the photometric variability and scene complexity commonly found in clinical videos.

— **The integration of deep learning-based enhancement and depth prediction modules into a passive SfM reconstruction pipeline**, allowing for the generation of geometrically consistent and spatially extended 3D cartography of hollow organ surfaces, such as the bladder and stomach.

— **Conducting a rigorous evaluation of the proposed framework** using both synthetic and patient endoscopic datasets. This includes quantitative metrics (e.g., depth accuracy, pose error) and qualitative assessments (e.g., surface coherence, visualization fidelity) to validate the clinical relevance and robustness of the method.

1.5.2 Medical Objectives of the Thesis ◂

The clinical motivation driving this thesis is to enhance the diagnosis, follow-up, and documentation capabilities of endoscopic procedures through advanced 3D reconstruction tools that remain

compatible with standard medical hardware and workflows. Conventional endoscopy relies heavily on real-time visual inspection, which limits post-hoc interpretation, precise spatial localization, and longitudinal tracking of pathological findings. This work aims to bridge that gap by introducing passive, image-based 3D reconstruction techniques that can be integrated into existing clinical routines without modifying current endoscopic systems.

The medical objectives pursued in this thesis are as follows:

— **Improvement of the visualization of lesions and anatomical landmarks** by generating extended mosaics of surfaces that offer a unified and spatially coherent view of pathological areas, aiding in a more comprehensive diagnosis.

— **Facilitating the longitudinal follow-up of patients** by enabling spatially comparisons between successive examinations. This allows clinicians to monitor lesion progression or healing across time with improved precision.

— **Allowing for data documentation and traceability** through the creation of 3D mosaics that can be archived and revisited. These reconstructions enable better record-keeping, multidisciplinary review, and medico-legal documentation.

— **Ensuring compatibility with real-world clinical constraints** by relying solely on passive image acquisition without any hardware modifications. This guarantees non-invasiveness and allows seamless integration into existing endoscopic platforms, avoiding the need for active vision systems or specialized lighting hardware.

1.6 Conclusion ◂

This chapter provided a comprehensive overview of the clinical and scientific motivations driving the development of advanced 3D reconstruction techniques in endoscopy. It critically examined the limitations of traditional monocular imaging systems—particularly their inability to offer spatial continuity, reproducibility, and robust post-operative assessment. The discussion underscored the clinical value of extended surface representations in enabling more precise lesion visualization, longitudinal follow-up, and enhanced data traceability.

A review of existing 3D reconstruction and mosaicing methods, from classical geometric techniques to emerging deep learning paradigms, highlighted the unique challenges of endoscopic imaging environments, including tissue deformation, specular reflections, and photometric instability. These observations motivated the exploration of hybrid strategies that combine passive vision methods with deep learning-based enhancements. The chapter concluded by formally articulating the scientific and medical objectives of the thesis, laying the foundation for the proposed methodological contributions.

Chapter 2

Deep Learning-based Image Enhancement in Endoscopy

2.1 Introduction ◂

Endoscopic imaging has become an indispensable tool in minimally invasive diagnosis and therapeutic procedures, particularly within the gastrointestinal (GI) tract. Despite significant improvements in optical design and imaging electronics, endoscopic video sequences frequently exhibit illumination-related artifacts, including, notably, (i) vignetting due to the optics of the endoscopes, (ii) specular reflections, and (iii) local over- and/or underexposures of tissues occurring either individually or simultaneously in several parts of the image.

As shown in Fig. 2.1, local under- or overexposures typically appear on inner tissues having particular local shapes and orientations with respect to the endoscope's optical axis. The under- or over-exposition severity also relates to the proximity of the mucosa to the endoscope's distal tip from which the scene illumination light is emanating (Ali et al. [2020]). In endoscopy, such unwanted exposure effects not only degrade the visual perception but also compromise the performance of downstream computer-assisted methods such as lesion detection (Ali et al. [2021b]), tissue (Zhao et al. [2021]) or object (Martínez et al. [2020]) classification, and three-dimensional (3D) reconstruction (Ma et al. [2019]).

FIGURE 2.1 – Frames extracted from a colonoscopy video demonstrate how camera motion influences dynamic exposure changes. As the endoscope advances through the gastrointestinal tract, lighting conditions transition from evenly illuminated scenes to regions of localized overexposure, primarily caused by close proximity to the mucosal surface. These abrupt variations violate the photometric consistency assumptions often required by subsequent computer vision algorithms.

Image enhancement (IE) methods have been extensively investigated as a means to mitigate the limitations involved by photometric artifacts in endoscopic imaging. Conventional approaches—such as histogram equalization, gamma correction, or Retinex-based decomposition—provide only partial improvements, as they typically address either global or local exposure issues, but rarely both simultaneously. These methods often fail to generalize across diverse illumination conditions and struggle to preserve fine anatomical structures, particularly in cases where underexposed and overexposed regions coexist within the same frame (Guo et al. [2017b]; Zhang et al. [2019a]). Furthermore, many of these techniques involve computationally intensive operations that hinder their application in real-time clinical environments, where latency and frame rate are critical constraints.

Recent advances in deep learning (DL) have enabled a new class of data-driven IE methods capable of learning complex mappings between corrupted and uncorrupted visual domains. Although these approaches have shown promising results in non-medical domains (Afifi et al. [2021]), their translation to endoscopy is impeded by two main factors: (i) the lack of paired datasets (i.e., same or similar images with and without exposure artifacts) designed specifically for endoscopy, and (ii) the absence of robust and unified models that can enhance images affected by both types of exposure errors in real time.

This thesis introduces the *Endo4IE* dataset —a synthetic, reference-based benchmark constructed from annotated colonoscopic video sequences— to address the lack of paired, photometrically degraded and reference-quality endoscopic images for supervised training. The dataset is designed to simulate various clinically realistic under- and overexposure through controlled image-to-image translation, which provides aligned pairs of degraded frames and frames without photometric degradations suitable for training and evaluating enhancement models. Based on this benchmark, two original deep learning-based methods (Endo-LMSPEC and Endo-ViT) are proposed to improve image quality in endoscopic video sequences and to support robust downstream tasks such as monocular depth estimation and 3D reconstruction.

This chapter addresses the core challenges in exposure correction for endoscopic imaging. It begins by analyzing the photometric artifacts specific to this imaging modality. Next, it describes

the construction and structure of the Endo-4IE dataset, which serves as a foundation for training and benchmarking supervised IE models. It then details two contributions to state-of-the-art learning-based image enhancement approaches: Endo-LMSPEC and Endo-ViT. Finally, an evaluation and comparison of their performance is both quantitatively and qualitatively given using standard metrics and with a discussion of their practical advantages for downstream medical imaging tasks.

2.2 Image Enhancement: Principles and Evaluation Criteria ◂

IE techniques aim to improve the visual quality of an image while preserving its inherent structural and spatial characteristics. In particular, the scale, resolution, and diagnostic content of the image must remain unaltered. Common degradations that enhancement methods seek to address include low resolution, poor contrast, under- or overexposure, saturation imbalances, and noise corruption. Formally, an image enhancement transformation can be represented as a function $\mathscr{T}$ mapping an input image $A(x,y)$ to an enhanced output $B(x,y)$, as given in Eq. (2.1):

$$B(x,y) = \mathscr{T}[A(x,y)] \tag{2.1}$$

Figure 2.2 illustrates several elementary enhancement operations, such as brightness and contrast modulation or image negation. These operations are implemented through pixel-wise transformations commonly encountered in classical image processing approaches. (Jawdekar and Joshi [2021]).

The brightness modifications illustrated in Fig. 2.2 can be mathematically expressed by the following equations. The parameters k_{add}, k_{subs}, and k_{scale} control the intensity adjustment and are typically selected based on the dynamic range and content characteristics of the input image:

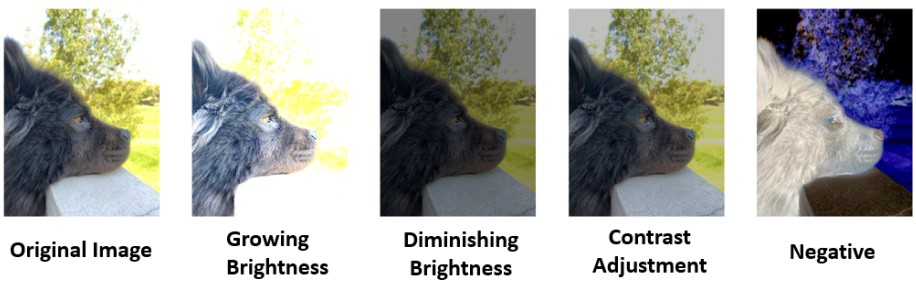


FIGURE 2.2 – Representative examples of intensity-based transformations applied for image enhancement. These transformations modify brightness, contrast, or inversion (negative) while preserving the spatial structure of the image (Jawdekar and Joshi [2021]).

Brightness increase:	$B(x,y) = A(x,y) + k_{add}$
Brightness reduction:	$B(x,y) = A(x,y) - k_{subs}$
Contrast scaling:	$B(x,y) = A(x,y) \cdot k_{scale}$
Negative transformation:	$B(x,y) = 255 - A(x,y)$

2.2.1 Evaluation Metrics for Image Enhancement ◂

The evaluation of the performance of an image enhancement model is commonly carried out using reference-based or non-reference-based image quality metrics. When paired ground truth is available, reference-based metrics are preferred due to their stronger correlation with perceptual fidelity.

The most widely used reference-based metrics (Wang et al. [2004]) are the Peak Signal-to-Noise Ratio (PSNR, see Eq. (2.18)) and the Structural Similarity Index Measure (SSIM, see Eq. (2.19)). These metrics evaluate the similarity between the enhanced image Y and a ground-truth image G, either in terms of pixel-level fidelity (PSNR) or in terms of perceived structural coherence (SSIM).

In Eq. (2.2), MAX_I denotes the maximum possible pixel intensity value in the image, defining the dynamic range $[0, MAX_I]$. For standard 8-bit images, this value is typically 255. The PSNR is computed based on the mean squared error (MSE) defined in Eq. (2.3), and is expressed in decibels (dB) to quantify the fidelity of a distorted image relative to its reference.

$$\text{PSNR} = 10 \cdot \log_{10}\left(\frac{MAX_I^2}{\text{MSE}}\right), \tag{2.2}$$

$$\text{MSE} = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[G(i,j) - Y(i,j)]^2, \tag{2.3}$$

In Eq. (2.19), the Structural Similarity Index Measure (SSIM) is computed over local image patches (typically of size 11×11), rather than entire images, to capture localized structural distortions. Let x and y be two corresponding patches from the reference and test images, respectively. The SSIM is then defined as:

$$\text{SSIM}(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}, \tag{2.4}$$

where:

— μ_x and μ_y are the mean intensities of patches x and y,

— σ_x^2 and σ_y^2 are the variances,

— σ_{xy} is the covariance between x and y,

— $c_1 = (k_1 MAX_I)^2$ and $c_2 = (k_2 MAX_I)^2$ are two small constants used to stabilize the division when the denominator is close to zero.

Standard values are $k_1 = 0.01$ and $k_2 = 0.03$. The overall SSIM index for an image is obtained by averaging the SSIM values computed over all patch locations.

In many medical imaging applications, reference-based evaluation is impractical due to the unavailability of paired data. Acquiring identical anatomical views under different controlled exposure conditions is often infeasible due to patient motion, procedural variability, and equipment constraints. In such contexts, no-reference (NR) image quality assessment (IQA) metrics offer a viable alternative for evaluating enhancement performance.

The most widely used NR-IQA metrics include:

— **Naturalness Image Quality Evaluator** (NIQE, (Mittal et al. [2012])): This metric evaluates the "unnaturalness" of an image by measuring how far its local statistics deviate from those observed in high-quality natural images. The process involves extracting statistical features—such as mean-subtracted contrast-normalized (MSCN) coefficients and products of neighboring MSCN coefficients—from spatial patches of the image. These features are modeled as a multivariate Gaussian (MVG) distribution

$$\text{NIQE}(I) = \sqrt{(\mu_t - \mu_n)^T \left(\frac{\Sigma_t + \Sigma_n}{2}\right)^{-1} (\mu_t - \mu_n)},$$

where (μ_t, Σ_t) are the mean and covariance of the test image features, and (μ_n, Σ_n) are the parameters of the MVG model trained on natural images. Lower NIQE values indicate better perceptual quality.

— **Ma Score** (Ma et al. [2017]): The Ma score is learned from aesthetic image judgments provided by human raters. A deep neural network (based on GoogLeNet) is trained to regress aesthetic scores from the AVA dataset. Let $f(I) \in [0, 10]$ be the predicted score for the image I. There is no explicit analytical formula for the Ma score, as it is the output of a neural network, but the score reflects perceptual preference. Higher values indicate greater visual appeal.

— **Perceptual Index** (PI, (Blau and Michaeli [2018])): The PI is a combined metric designed to balance handcrafted (NIQE) and learned (Ma) quality predictions. It is defined as:

$$\text{PI}(I) = \frac{1}{2}\left((10 - f(I)) + \text{NIQE}(I)\right),$$

where $f(I)$ is the Ma score and $\text{NIQE}(I)$ is the NIQE value for image I. A lower PI corresponds to better perceptual quality, as it reflects both low statistical deviation and high aesthetic preference.

These no-reference metrics are grounded in large datasets of aesthetically rated images, primarily sourced from photographic quality benchmarks. Although they were originally developed for natural scenes, they offer a practical proxy for evaluating photometric realism and structural plausibility in medical image enhancement, particularly when reference images are unavailable.

2.2.2 Image Enhancement Methods ◄

Image enhancement (IE) methods aim to improve the perceptual quality of an image while preserving its underlying structural and geometric integrity. Geometric preservation refers to the maintenance of spatial relationships, edge continuity, and the consistency in contours and surfaces of the image. This is particularly critical in medical imaging, where visual artifacts can mislead diagnostic interpretation.

This section reviews three primary families of IE techniques: (i) histogram-based methods, which adjust intensity distributions to enhance contrast; (ii) Retinex-based approaches, inspired by human visual perception of illumination and reflectance; and (iii) deep learning-based strategies, which learn complex mappings for illumination correction and texture restoration from large-scale datasets.

2.2.2.1 Histogram-Based Methods ◄

Histogram-based techniques adjust contrast by redistributing the intensity values of image pixels. These methods are particularly suited for images of low contrast in which the dynamic range of possible values is under-exploited.

— **Histogram Equalization (HE):** HE changes the pixel intensity histogram of an image so that it spans the full dynamic range. As shown in Fig. 2.3, HE enhances global contrast by equalizing the distribution of grayscale levels, though it may cause over-enhancement or amplify noise in homogeneous regions.

— **Adaptive Histogram Equalization (AHE):** AHE improves upon HE by applying contrast enhancement locally in small, overlapping image regions. This approach increases local contrast and can lead to finer details. However, it can also introduce noise in uniform areas. A common variant, Contrast Limited AHE (CLAHE), mitigates this risk by clipping histogram peaks.

— **Bi-Histogram Equalization (BBHE):** BBHE divides the intensity histogram into two sub-histograms at the mean gray-level of the image. Histogram equalization (HE) is then applied independently to each intensity interval—[0, mean] and [mean + 1, 255]—ensuring that the overall brightness is better preserved. This dual-interval strategy reduces the risk of over-enhancement and contrast saturation commonly associated with global HE.

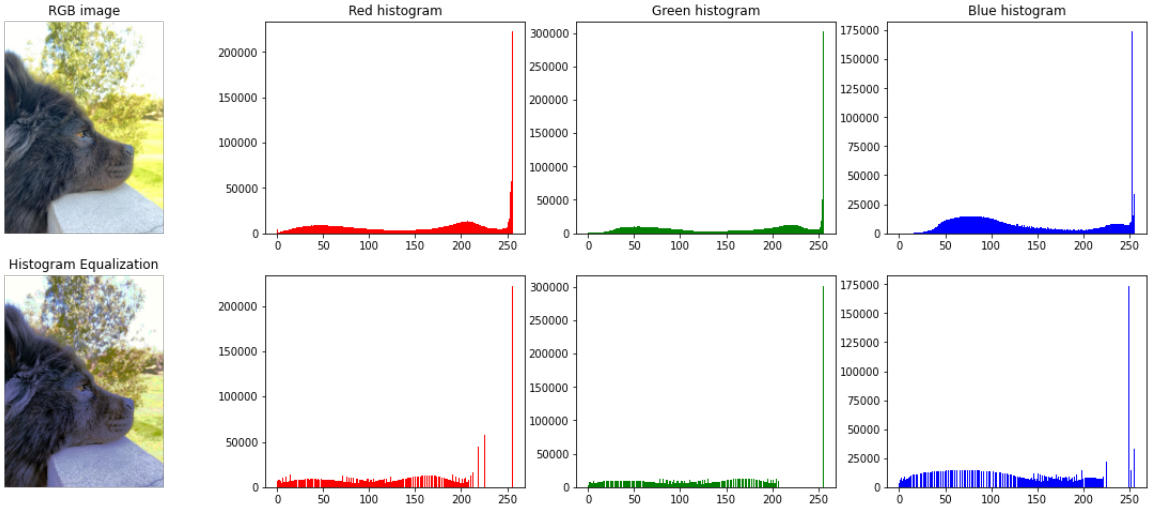


FIGURE 2.3 – Effect of histogram equalization on an RGB image and its corresponding color histograms. The first row presents the original image alongside its red, green, and blue channel histograms. The second row displays the enhanced image after applying histogram equalization, along with the redistributed histograms for each color channel, illustrating improved contrast and intensity spread Jawdekar and Joshi [2021].

2.2.2.2 Retinex Theory-Based Methods ◄

The Retinex theory, originally introduced by Land and McCann (Land and McCann [1971]), exploits the fact that an image $S(x,y)$ can be represented as the product of its reflectance $R(x,y)$ and illumination $I(x,y)$:

$$S(x,y) = R(x,y)I(x,y) \tag{2.5}$$

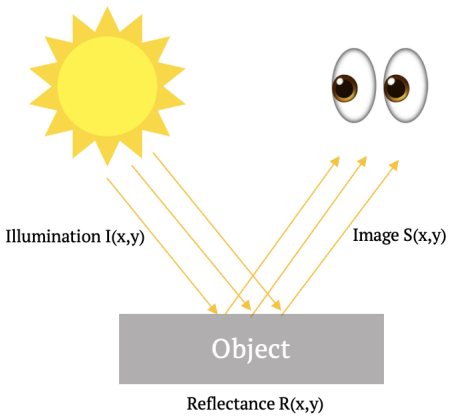


FIGURE 2.4 – Illustration of the image formation process: the observed image $S(x,y)$ is modeled as the product of the illumination $I(x,y)$ and the reflectance $R(x,y)$. This model is fundamental in separating lighting effects from inherent surface properties in image enhancement techniques Anoop and Deivanathan [2024].

Two notable methods based on Eq. (2.5) and the physical property in Fig. 2.4 are:

— **Low-Light Image Enhancement (LIME)** Guo et al. [2017b]: LIME estimates the illumination map of an image at the pixel level by taking the maximum value across the RGB channels for each pixel, i.e., $T(x,y) = \max\{R(x,y), G(x,y), B(x,y)\}$. This per-pixel illumination map is then refined using structure-aware smoothing techniques to guide the enhancement of underexposed regions. Although effective for improving visibility in dark areas, LIME is not designed to correct overexposed regions, and may fail in scenes with mixed illumination.

— The **DUAL Illumination Estimation** (Zhang et al. [2019b]) that addresses both under- and overexposure by computing a standard illumination map from the original image and a reverse map from the image negative. The combined output enables more comprehensive exposure correction.

2.2.2.3 Deep Learning-Based Methods ◄

Recent advancements in convolutional neural networks (CNNs) and generative adversarial networks (GANs) have led to the development of powerful data-driven image enhancement (IE) techniques. These models are capable of learning complex mappings from degraded to enhanced images by leveraging large-scale datasets.

The growing interest in deep learning-based exposure correction is illustrated in Fig. 2.5, which highlights representative scenarios of over- and underexposure. Early notable contributions in this domain include:

— **DPED (DSLR Photo Enhancement Dataset)** (Ignatov et al. [2017]): This method uses a supervised learning approach to enhance low-quality images captured by smartphone cameras, translating them to DSLR-like quality. The network is trained on a dataset of paired smartphone and DSLR photos of the same scene, aligning low- and high-quality views at pixel level.

— **DPE (Deep Photo Enhancer)** (Chen et al. [2018b]): DPE is a GAN-based framework designed for generic photographic enhancement. It combines a residual learning architecture with perceptual and adversarial losses to correct global tone, contrast, and texture, improving the overall aesthetic quality of images without requiring paired high-quality references.

— **RetinexNet** (Wei et al. [2018]) and **KinD (Knowledge-inspired Deep Retinex Decomposition)** (Zhong et al. [2020]): Both methods are grounded in Retinex theory, which models an image as the product of reflectance and illumination components. RetinexNet performs joint decomposition and enhancement, while KinD introduces knowledge-based constraints to improve decomposition accuracy and enable robust enhancement of underexposed regions.

— **Zero-DCE (Zero-Reference Deep Curve Estimation)** (Guo et al. [2020]) and **EnlightenGAN** (Jiang et al. [2021b]): These methods enable low-light enhancement in the absence

of paired training data. Zero-DCE formulates enhancement as a task of learning pixel-wise light adjustment curves under a self-supervised loss, while EnlightenGAN uses an unpaired image-to-image translation framework to learn exposure corrections with perceptual and color consistency constraints.

Recent work by Afifi et al. (Afifi et al. [2021]) and Lv et al. (Lv et al. [2020]) addresses both under- and overexposure within a unified architecture, a requirement for complex environments such as endoscopy.

2.2.2.4 Learning Multi-Scale Photo Exposure Correction (LMSPEC) ◄

LMSPEC (Afifi et al. [2021]) is a deep convolutional model designed to correct both under- and overexposed image regions by exploiting multi-scale representations. The LMSPEC-pipeline is given in Fig. 2.6.

The model processes the input image I by extracting n random patches $I'_1, \ldots, I'_n$, which are decomposed into Laplacian pyramids (LPs). Ground-truth patches undergo Gaussian pyramid decomposition. Each LP level is processed by a U-Net-like subnetwork in a cascaded configuration, enabling fine-grained correction of structural and photometric details.

Training proceeds in three stages, with patches of increasing size (128×128, 256×256, and 512×512 pixels). The final model contains approximately 7 million parameters and is optimized using a composite loss:

$$\mathscr{L} = \mathscr{L}_{\text{pyr}} + \mathscr{L}_{\text{rec}} + \mathscr{L}_{\text{adv.}} \tag{2.6}$$

The three complementary components of the composite objective function defined in Eq. (2.6) guide the training process of the proposed image enhancement network to capture both pixel-level fidelity and perceptual quality:

— *Pyramidal loss* $\mathscr{L}_{\textbf{pyr}}$ is designed to enforce multiscale consistency between the enhanced out-

FIGURE 2.5 – Low-light enhancement examples using deep learning methods (Ignatov et al. [2017]; Chen et al. [2018a]; Wei et al. [2018]; Jiang et al. [2021b]; Guo et al. [2020]).

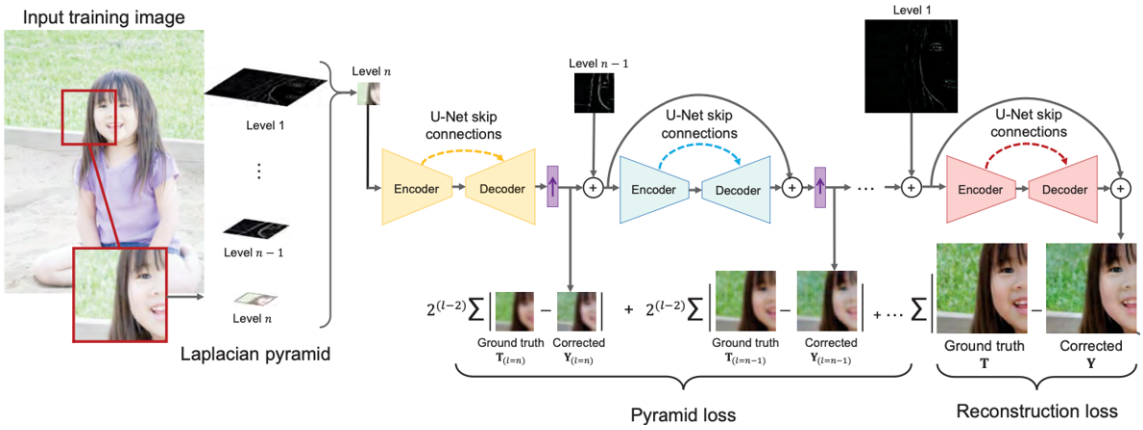


FIGURE 2.6 – Overview of the training framework described in (Afifi et al. [2021]). It combines Laplacian pyramid decomposition, U-Net-based reconstruction, and composite loss functions. The input image is first decomposed into multiple frequency levels with the Laplacian pyramid. At each level, a U-Net model with skip connections performs enhancement, guided by the pyramid loss. The outputs across all levels are combined to compute the final reconstruction, which is compared to the ground truth using a reconstruction loss. Additionally, an adversarial discriminator trained with a sigmoid cross-entropy loss enforces perceptual realism (Afifi et al. [2021]).

put and the ground truth. By comparing image representations across different spatial resolutions, this term ensures structural coherence at both coarse and fine levels, which is critical for preserving textures and semantic integrity.

— *Reconstruction loss* $\mathscr{L}_{\mathbf{rec}}$ is determined with a ℓ_1 or ℓ_2 norm of the pixel value differences between the predicted image and the reference (ground truth). This term penalizes pixel-wise deviations and encourages accurate recovery of intensity values, thereby guiding the network to produce outputs faithful to the input distribution.

— *Adversarial loss* $\mathscr{L}_{\mathbf{adv}}$ is incorporated following the generative adversarial network (GAN) framework. It aims to push the output distribution closer to the distribution of real images by encouraging the generator to produce perceptually plausible results. This term is essential for enhancing the natural appearance of the enhanced images.

Total loss $\mathscr{L}$ is thus a balanced combination of perceptual, content, and adversarial constraints, promoting both quantitative accuracy and visual realism in the enhanced images.

2.2.3 Standard Datasets ◂

Benchmarking image enhancement methods requires an access to high-quality, diverse, and well-annotated datasets. In the context of natural images, several widely used benchmarks exist, including the LOw-Light (LOL) dataset (Wei et al. [2018]), the Converted See-In-the-Dark (CSID) dataset (Chen et al. [2018a]), and the Adobe FiveK dataset (Bychkovsky et al. [2011]). Additionally,

domain-specific datasets such as the Berkeley Deep Driving (BDD-100K, (Yu et al. [2020])) focus exclusively on low-light conditions in street scenes, primarily for autonomous driving applications.

In contrast, medical imaging lacks standardized datasets specifically curated for image enhancement tasks. This gap persists despite the clinical relevance of high-quality image data. Several recent efforts have attempted to overcome this limitation by generating synthetic datasets. For instance, synthetic MRI-to-CT translation frameworks have been proposed to simulate paired datasets, although concerns regarding distribution shifts and model generalizability have been raised (Cohen et al. [2018]). Other studies have adopted data augmentation strategies such as downsampling and upsampling to simulate degradation, particularly in low-resolution and blur correction tasks (Almalioglu et al. [2020b]).

This shortage of medically validated, paired datasets for exposure correction underscores the importance of synthetic but anatomically grounded benchmarks such as the *Endo4IE* dataset introduced in this thesis. By providing reference-based pairs of underexposed, overexposed, and photometrically corrected endoscopic images, the dataset aims to support the development and evaluation of learning-based IE methods in clinically realistic settings.

2.3 New Exposure Correction Method ◂

2.3.1 Development of the Endo-4IE Dataset ◂

The supervised training of deep learning-based photometric enhancement models is hindered by the absence of paired endoscopic images exhibiting both degraded and reference-quality illumination. In clinical conditions, it is impossible to acquire a large set of images of same anatomical regions seen under different exposures (it is tedious even for experimented endoscopists and not enough time is available during the examinations). As a result, existing endoscopic datasets lack the ground-truth supervision typically required for training modern enhancement architectures.

The **Endo-4IE** dataset was developed as a *synthetic reference-based benchmark* specifically designed for photometric enhancement tasks in endoscopy. The term "synthetic" refers to the fact that degraded images exhibiting under- and overexposure artifacts are not acquired through real-world variations, but are artificially generated from clinically acquired reference frames using image-to-image translation techniques. The construction of this dataset follows a pipeline that incorporates object detection, domain-specific content filtering, and generative modeling via CycleGANs. This pipeline ensures that only diagnostically relevant frames are used and that corresponding degraded and reference images are spatially and anatomically aligned. As a result, **Endo-4IE** provides a reproducible and task-specific benchmark for training and evaluating deep learning-based enhancement models under controlled photometric degradation scenarios.

The dataset is designed to support supervised training and evaluation of models aimed at correc-

ting both underexposure and overexposure in endoscopic imaging, and it provides sufficient photometric diversity to test generalization under realistic lighting fluctuations. It comprises three types of images: well-exposed reference frames, synthetically overexposed frames (i.e., overexposed frame parts up to images with local saturation), and synthetically underexposed frames characterized by global dimming and contrast loss.

Although several publicly available datasets exist for endoscopic tasks such as classification Borgli et al. [2020]; Jiang et al. [2021a], segmentation Bernal et al. [2018], lesion detection Ali et al. [2020, 2021a], and even 3D reconstruction Ma et al. [2022]; Mahmood et al. [2022], there is currently no dedicated benchmark specifically designed for photometric image enhancement. Most of these datasets are curated to support structural or semantic interpretation tasks and assume a relatively stable illumination profile across frames. Consequently, they do not systematically capture or annotate common photometric degradations such as overexposure, underexposure, or specular reflections—despite the fact that such artifacts are frequently encountered in real-world clinical settings and significantly impact visual quality and algorithm performance.

In this work, we leverage three such repositories—EAD2020 Ali et al. [2020], EAD2.0 Ali et al. [2021a], and HyperKvasir Borgli et al. [2020]—to assemble a dataset tailored for exposure correction. A rigorous quality control step was first applied to discard non-diagnostic frames, including those that were entirely dark, oversaturated, or motion-blurred, yielding a curated collection of 7,064 high-quality images. These images were then automatically categorized according to their photometric properties using a YOLOv4-based detector trained on expert-labeled examples from EAD2020. The classification scheme divided the dataset into three categories: (i) well-exposed frames, (ii) underexposed regions (typically caused by poor lighting geometry or occlusion) , (iii) overexposed regions (often due to proximity-induced saturation or specular reflections). These categories were subsequently used to construct paired samples for supervised learning.

Frames identified as well-exposed were selected as reference images. A CycleGAN architecture (Zhu et al. [2017]) was then trained to generate corresponding underexposed and overexposed versions of these images. Unlike supervised GANs, CycleGAN does not require paired input-output examples and is therefore suitable for this domain, where only unpaired exemplars exist. The translation models were trained using 1,200 real underexposed and 1,200 real overexposed frames as target domains. During the translation, each well-exposed image was passed through both generator models to produce the two corresponding corrupted variants.

As shown in the pipeline overview given in Fig. 2.7, a filtering step based on the Mean Squared Error (MSE) and Structural Similarity Index (SSIM) metrics was performed to select images, which allows ensuring the quality and photometric diversity of the generated samples.

The quality of a generated sample in this context refers to the preservation of structural and textural features essential for meaningful downstream processing, while allowing for sufficient photometric variation to enrich training diversity. The SSIM metric was used to evaluate the pre-

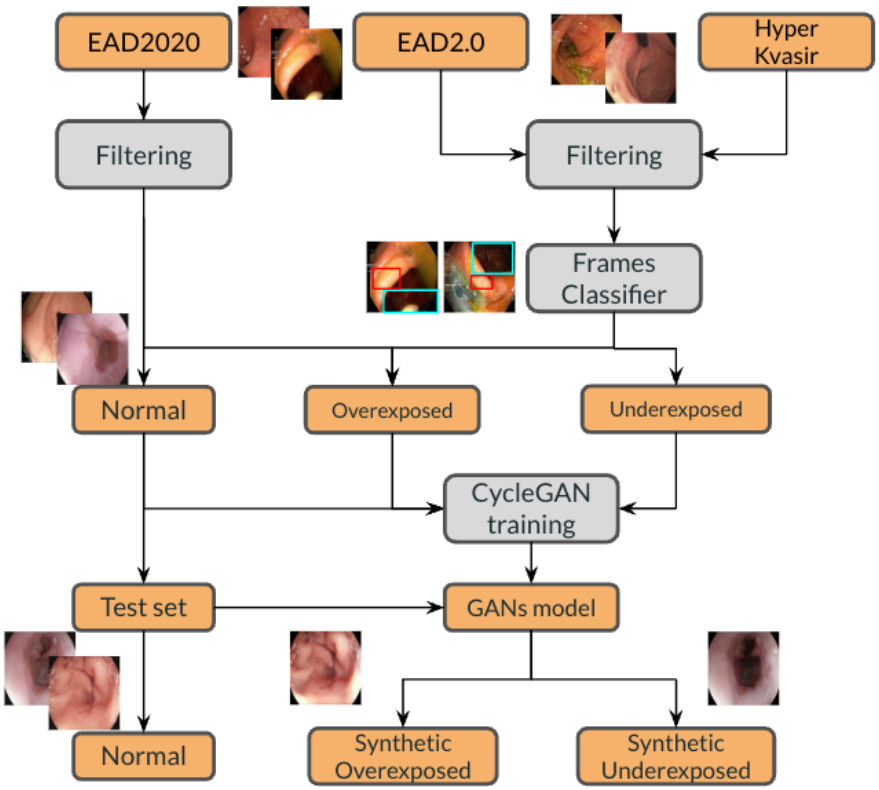


FIGURE 2.7 – Overview of the **Endo-4IE** dataset construction pipeline. An object detector trained on expert-labeled frames identifies photometric categories in unlabeled endoscopic data. Reference-quality frames are translated into underexposed and overexposed domains via CycleGANs. The resulting image triplets are filtered using MSE and SSIM constraints to preserve semantic content (textures and structures) while introducing local under- and over-exposures (Vega et al. [2022]).

servation of structural integrity, ensuring that essential spatial configurations—such as edges and contours—remain intact. Conversely, MSE quantifies the pixel-wise intensity differences, capturing photometric deviations such as changes in illumination or contrast.

Only synthetic images that both exhibit sufficient intensity deviation from their original (MSE measurement), and retain their structural coherence (assessed with the SSIM) were retained. Specifically, only samples with an MSE between 100 and 1500 and an SSIM in the range of $[0.6, 0.9[$ were selected. This dual-threshold approach guarantees that selected images are neither trivial (too similar) nor unusable (overly degraded), thus balancing novelty and usability in the augmented dataset.

The final **Endo4IE** dataset includes 2,216 simulated images comprising 1,231 synthetically overexposed images and 985 synthetically underexposed images. All frames were resized to 512×512 pixels and aligned to form 985 paired triplets, each consisting of a reference image and its corresponding overexposed and underexposed versions. These triplets were partitioned into training (70%), testing (27%), and validation (3%) subsets, ensuring diversity in anatomical content and exposure conditions across splits.

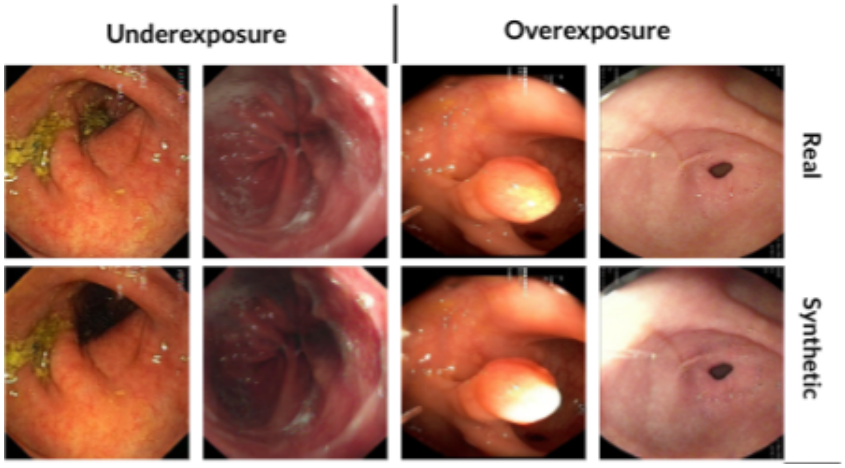


FIGURE 2.8 – Examples of real and synthetic exposure artifacts in colonoscopic images from the **Endo4IE** dataset. The top row illustrates real clinical frames exhibiting underexposure (left) and overexposure (right). The bottom illustrates synthetically generated counterparts using image-to-image translation, preserving anatomical realism while simulating non-optimal exposure. These paired examples enable supervised training and evaluation of deep learning-based exposure correction models.

Figure 2.8 illustrates representative frames from the dataset, including reference images and their synthetically corrupted counterparts. The examples highlight the variability introduced during GAN-based synthesis and underscore the value of the dataset for benchmarking both traditional and learning-based correction methods. The Endo4IE dataset is publicly available through the Mendeley Data repository (Vega et al. [2022]) and serves as a cornerstone for reproducible evaluation in photometric correction tasks within medical image analysis.

Proposed Image Enhancement Methods for Endoscopy

For complex medical image modalities such as ultrasound and endoscopy, classical IE methods often fall short. In such cases, deep learning-based IE methods have shown superior performance. The authors in (Miao et al. [2021]) applied a convolutional neural network to enhance ultrasound images of patients with ureteral calculi. Similarly, in the endoscopy, tasks such as resolution enhancement, blur correction, and restoration of occluded regions are increasingly addressed using learning-based methods (Almalioglu et al. [2020a]; Tao et al. [2018]; Kohler et al. [2014]). These models not only offer improved visual quality but also enhance the reliability of downstream tasks such as lesion detection, segmentation, and 3D reconstruction.

Sections 2.3.2 and 2.3.3 present two DL-based IE-methods developed in the frame of this thesis, namely the Endo-LMSPEC architecture, and the Endo-ViT model, a color-aware transformer-based solution for real-time enhancement.

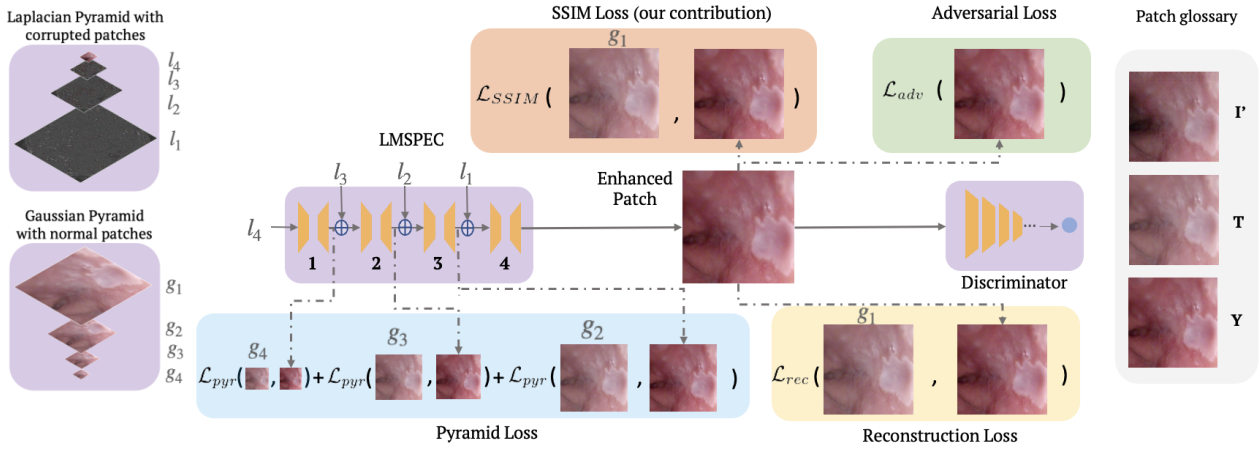


FIGURE 2.9 – Architectural overview of the proposed Endo-LMSPEC model for endoscopic image enhancement. The method leverages a Laplacian pyramid decomposition of corrupted input patches, with each frequency band processed by a dedicated multi-scale U-Net sub-network. The outputs are combined and refined under a composite loss function comprising pyramid loss, reconstruction loss, SSIM-based structural preservation, and adversarial loss. The patch glossary illustrates the input corrupted patch (I'), the ground truth reference patch (T), and the enhanced output patch (Y).

2.3.2 Endo-LMSPEC, a Multi-Scale Exposure Correction Network ◂

The Endo-LMSPEC model (Vega et al. [2023]) extends the LMSPEC architecture (Afifi et al. [2021]) to address the challenges of photometric inconsistency in endoscopic imaging. As sketched in Fig. 2.9, it is designed to handle exposure correction across multiple spatial frequencies by leveraging a Laplacian pyramid decomposition of the input image. Each frequency band is independently processed using dedicated U-Net-style sub-networks, enabling coarse-to-fine corrections—from global illumination shifts to fine-grained textural enhancements. The final output is reconstructed by summing the enhanced pyramid levels.

Patch-Based Learning. The network is trained on randomly sampled image patches of size 128×128 and 256×256 pixels to manage computational complexity and focus on localized artifacts. This approach enables fine detail modeling while reducing memory requirements. However, during inference, this strategy may introduce *tiling artifacts*, particularly in high-frequency or non-uniform lighting regions. These artifacts manifest as seams along patch boundaries if proper blending or overlap is not employed. Furthermore, the separate processing of pyramid levels introduces additional computational overhead, resulting in a model with approximately 7 million parameters and an inference speed of only 8 FPS—rendering it less suited for real-time clinical use.

Loss Formulation. A composite loss ($\mathscr{L}_{\text{total}}$ in Eq. (2.7)) is used to effectively train the model. It combines photometric, structural, perceptual, and adversarial objectives.

$$\mathscr{L}_{\text{total}} = \alpha\mathscr{L}_{\text{pyr}} + \beta\mathscr{L}_{\text{rec}} + \gamma\mathscr{L}_{\text{SSIM}} + \delta\mathscr{L}_{\text{adv}}, \tag{2.7}$$

where $\alpha, \beta, \gamma, \delta \in \mathbb{R}^+$ are positive scalar weights empirically determined through cross-validation, as detailed in Section 2.4.3.

Pyramid Loss $\mathscr{L}_{\text{pyr}}$. This term ensures consistency across frequency bands. For each Laplacian pyramid level $i = 2, 3, 4$, the model output $\hat{l}_i$ is compared to the corresponding ground truth Gaussian level g_i as:

$$\mathscr{L}_{\text{pyr}} = \sum_{i=2}^{4} \left(2^{i-2} \cdot \frac{1}{N_x^i \times N_y^i} \sum_{x=1}^{N_x^i} \sum_{y=1}^{N_y^i} \left| \hat{l}_i(x,y) - g_i(x,y) \right| \right), \tag{2.8}$$

where $N_x^i \times N_y^i$ denotes the number of pixels (i.e., columns multiplied by rows) at the resolution of the image at Laplacian pyramid level i. The weighting factor 2^{i-2} assigns greater importance to finer-scale features, emphasizing high-frequency details critical for perceptual sharpness.

Reconstruction Loss $\mathscr{L}_{\text{rec}}$. An $\mathscr{L}_1$ norm is applied on the color differences between the final enhanced patch Y_j and its ground truth counterpart T_j to preserve global pixel-level fidelity:

$$\mathscr{L}_{\text{rec}} = \frac{1}{M \cdot N} \sum_{i=1}^{M} \sum_{j=1}^{N} \left| Y_{i,j} - T_{i,j} \right|, \tag{2.9}$$

In Eq. (2.9), N denotes the number of pixels per patch, and M represents the total number of patches extracted from the image. To accurately reflect the contribution of all patches during training, the reconstruction loss is computed as an average over both spatial and patch dimensions, ensuring that local deviations are properly aggregated at the image level.

Structural Similarity Loss ($\mathscr{L}_{\text{SSIM}}$). This component improves perceptual quality by evaluating luminance,

$$\mathscr{L}_{\text{SSIM}} = \frac{1}{M} \sum_{i=1}^{M} \frac{1 - \text{SSIM}(T_i, Y_i)}{2}, \tag{2.10}$$

In Eq. (2.10), M denotes the number of patches and W the number of local windows within which SSIM is evaluated within each patch. $\text{SSIM}(T_{i,j}, Y_{i,j})$ computes the structural similarity index between the predicted and ground-truth local windows. This formulation ensures that the loss captures both global consistency across patches and local perceptual quality within each region, thus promoting edge preservation and structural coherence across the entire image.

Adversarial Loss ($\mathscr{L}_{\mathbf{adv}}$). To enhance the perceptual realism of texture and fine structures, an adversarial discriminator D is trained to distinguish between real (ground truth) and enhanced (generated) image patches. The generator is penalized when its outputs are classified as fake. The adversarial loss used to update the generator is defined as:

$$\mathscr{L}_{\text{adv}} = -3hwn \cdot \log\left(S(D(Y_j))\right), \tag{2.11}$$

where:

— Y_j is the enhanced RGB patch at pyramid level j,

— $h \times w$ is the spatial resolution (height and width) of the patch,

— n is the number of pyramid levels used in multi-scale supervision,

— $S(\cdot)$ denotes the sigmoid activation function applied to the discriminator output.

The scalar multiplier $3hwn$ accounts for the total number of predicted values in the adversarial map: 3 channels (RGB), $h \times w$ spatial positions per patch, and n pyramid levels. This normalization ensures that the adversarial signal is proportionally weighted with respect to the number of pixels evaluated across all levels and channels.

Summary. By combining frequency-specific enhancement, pixel-level accuracy, perceptual quality, and adversarial realism, Endo-LMSPEC provides a comprehensive solution to exposure correction. Despite its computational cost, it demonstrates excellent performance on the Endo4IE dataset and remains a strong convolutional baseline against which more efficient transformer-based models can be evaluated.

2.3.3 Endo-ViT, a Color-Aware Transformer-Based Enhancement ◂

The limitations of convolutional architectures like Endo-LMSPEC relate to their high parameter count, their limited receptive field, and their poor scalability for real-time deployment. The Endo-ViT model (Espinosa et al. [2024a]) was introduced as a lightweight transformer-based alternative for endoscopic image enhancement to take these limitations into account. Derived from the Illumination Adaptive Transformer (IAT, (Cui et al. [2022])), the Endo-ViT architecture is tailored to the photometric and structural demands of endoscopic imaging through the integration of three key design elements:

(i) **Color Normalization Layer:** To address color inconsistencies arising from heterogeneous illumination and variations in endoscopic imaging systems, a channel-wise learnable affine transformation is incorporated into each transformer block.

The transformation is defined as:

$$\text{Aff}(x) = \gamma x + \beta,$$

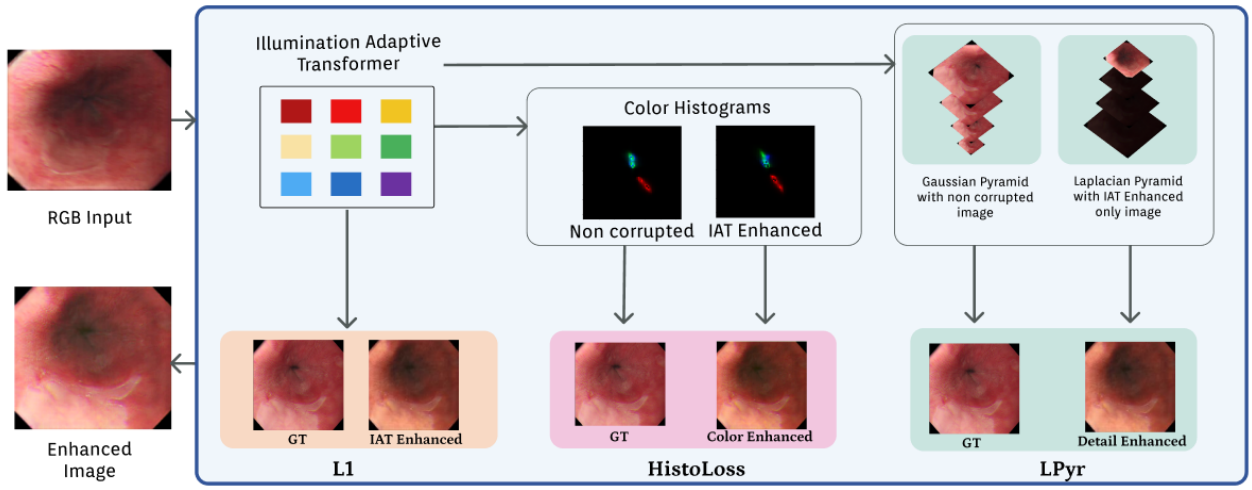


FIGURE 2.10 – Architectural overview of the Endo-ViT model. Endo-Vit integrates a local transformer branch to estimate spatial exposure corrections and a global branch that mimics ISP adjustments. The outputs are fused to form the enhanced image, supervised by a compound loss that enforces pixel, perceptual, and color distribution fidelity.

where:

— $x \in \mathbb{R}^{C\times H\times W}$ is the input feature map with C channels, height H, and width W,

— $\gamma, \beta \in \mathbb{R}^{C}$ are learnable vectors of scaling and bias parameters, respectively, applied independently to each channel,

The role of $\mathrm{Aff}(x)$ is to normalize the color distribution across channels in a content-aware and trainable manner. The multiplicative term γ adjusts the contrast or intensity per channel, while β shifting the mean, thus enabling the network to compensate for inter-channel imbalances without requiring hand-crafted color correction.

(ii) **Laplacian Pyramid Loss:** Inspired by Endo-LMSPEC (Vega et al. [2023]), A frequency-aware loss term that preserves edge and structural details by enforcing consistency across multiple spatial resolutions.

(iii) **Histogram Color Loss:** Inspired by HistoGAN (Afifi and Brown [2021]), this loss promotes chromatic consistency between enhanced outputs and ground truth references using histogram statistics in log-chroma space.

Architecture Overview. Endo-ViT is structured around two synergistic branches (see Fig. 2.10): - The **Local Branch** implements spatially-adaptive enhancement using transformer encoders equipped with Position-wise Enhancement Modules (PEMs) and depth-wise convolutions. These components enable local context modeling while maintaining computational efficiency. The inclusion of the aforementioned color normalization layer stabilizes hue and contrast variations in the image.
- The **Global ISP (Image Signal Processing) Branch** draws inspiration from traditional camera ISP pipelines, which are responsible for transforming raw sensor data into visually interpretable

images through a sequence of calibrated operations. These typically include white balance, gamma correction, color space conversion, and contrast enhancement—steps that emulate the behavior of real-world imaging hardware.

In the proposed architecture, this branch is designed to model such global photometric transformations in a learnable and endoscopy-specific manner. It predicts image-level correction parameters that mimic the effects of camera ISP modules, thereby compensating for illumination-induced distortions and sensor-specific artifacts. Specifically, given an input feature map X, the branch estimates a global affine transformation comprising a learned 3×3 color correction matrix M and a trainable gamma value γ. The output is computed as:

$$\hat{Y} = \Gamma(MX), \quad \text{with} \quad \Gamma(z) = z^{\gamma},$$

where $M \in \mathbb{R}^{3\times3}$ performs channel-wise color adaptation, and $\Gamma(\cdot)$ applies element-wise gamma correction. This formulation allows the model to correct global exposure and chromatic inconsistencies in a physically interpretable and computationally efficient manner.

$$\hat{Y} = \text{ISP}(X) = \Gamma(MX), \tag{2.12}$$

where $M \in \mathbb{R}^{3\times3}$ is a trainable color matrix, and $\Gamma(z) = z^{\gamma}$ is a per-channel gamma correction function $\gamma \in \mathbb{R}^{+}$ learned during training. This formulation allows for physically meaningful global color transformations.

The outputs of both branches are fused through element-wise summation, followed by a shallow convolutional refinement layer to yield the final enhanced image.

2.3.4 Loss Formulation for Endo-ViT ◂

The learning process is guided by a loss function $\mathscr{L}_{\text{total}}$ defined by:

$$\mathscr{L}_{\text{total}} = \mathscr{L}_1 + \lambda\,\mathscr{L}_{\text{pyr}} + \beta\,\mathscr{L}_{\text{hist}}, \tag{2.13}$$

where:

— $\mathscr{L}_1$ is the pixel-wise reconstruction loss,

— $\mathscr{L}_{\text{pyr}}$ enforces multi-scale detail preservation and

— $\mathscr{L}_{\text{hist}}$ ensures consistency in color distribution.

For the Endo-ViT model, the regularization parameters associated with the compound loss function given in Eq. (2.13) were selected based on extensive empirical evaluation. Specifically, the weight assigned to the Laplacian pyramid loss was set to $\lambda = 0.1$, while the histogram color loss was set to $\beta = 0.04$. These values were chosen to ensure a balanced trade-off between multi-scale

structural fidelity and chromatic consistency. The configuration was inspired by findings in the work of Cui et al. (Cui et al. [2022]), where a similar formulation was used for low-light image enhancement with an emphasis on brightness correction and detail preservation. Our own experimentation across validation folds confirmed that this weighting scheme yielded optimal visual quality and quantitative performance in endoscopic image enhancement tasks.

Pixel-wise $\mathscr{L}_1$ loss. This loss penalizes absolute intensity errors at each pixel:

$$\mathscr{L}_1 = \frac{1}{N}\sum_{j=1}^{N}\left|\hat{Y}_j - Y_j\right|, \tag{2.14}$$

where $\hat{Y}$ is the enhanced image, Y the ground truth, and N the total number of pixels. It serves as the backbone for minimizing low-level photometric discrepancy.

Laplacian Pyramid Loss $\mathscr{L}_{\text{pyr}}$. It captures structural differences across scales, both $\hat{Y}$ and Y being decomposed into Laplacian pyramids. The loss is computed as:

$$\mathscr{L}_{\text{pyr}} = \sum_{i=2}^{4} 2^{i-2}\cdot\frac{1}{N_i}\sum_{j=1}^{N_i}\left|\hat{L}_{i,j} - L_{i,j}\right|, \tag{2.15}$$

with $\hat{L}_{i,j}$ and $L_{i,j}$ denoting the Laplacian representations of predicted and ground truth images at level i and pixel j. N_i stands for the pixel number at that level i. The weights 2^{i-2} emphasize detail at finer scales.

Histogram Color Loss ($\mathscr{L}_{\text{hist}}$). To promote photometric realism and chromatic consistency, the model incorporates a histogram-based loss adapted from HistoGAN Afifi and Brown [2021]. This loss operates on 2D color histograms constructed in log-chroma space, where each histogram is normalized to sum to one. The histogram loss quantifies the statistical divergence between the enhanced and ground truth images using an approximation of the Hellinger distance. The final expression is given by:

$$\mathscr{L}_{\text{hist}} = \frac{1}{\sqrt{2}N}\sqrt{\sum_{n=1}^{N}\left(\sqrt{H(n)} - \sqrt{\hat{H}(n)}\right)^2}, \tag{2.16}$$

where:

— $\hat{H}(n)$ and $H(n)$ denote the values of the normalized 2D histograms for the enhanced and ground truth images at bin n,

— N is the total number of histogram bins,

— The square root operation is applied element-wise before computing the $\mathscr{L}_2$ norm.

This formulation ensures that subtle differences in color distribution are penalized while emphasizing discrepancies in sparse regions. The resulting value is always non-negative and differentiable, making it suitable for gradient-based optimization.

Summary. Endo-ViT unifies transformer-based modeling with a carefully engineered loss formulation that balances pixel accuracy, structural fidelity, and color realism. Its compact design and high performance make it a compelling alternative to CNN-based methods for enhancing endoscopic video in real time.

2.4 Performance Evaluation ◂

This section presents the evaluation of the image enhancement methods previously described, including traditional baselines mentioned in bibliographic Section 2.2.2, the Endo-LMSPEC model, and the transformer-based Endo-ViT architecture. The assessment includes both quantitative and qualitative analyses conducted on the Endo-4IE dataset. The experiments were designed to evaluate structural fidelity, photometric consistency, and real-time feasibility in the context of endoscopic image enhancement.

2.4.1 Evaluation Metrics ◂

Four evaluation metrics were employed to assess the quality of image enhancement models. Among them, three are full-reference metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Mean Squared Error (MSE). These metrics were computed by comparing enhanced frames against their corresponding paired ground truth images from the Endo-4IE dataset, which provides synthetically altered and unaltered versions of the same scenes.

The **Mean Squared Error (MSE)** quantifies the average squared intensity difference between the enhanced image Y and its reference T, over $m \times n$ pixels:

$$\mathrm{MSE} = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \left(Y_{i,j} - T_{i,j} \right)^2 . \tag{2.17}$$

The **Peak Signal-to-Noise Ratio (PSNR)** measures the logarithmic ratio between the maximum possible pixel intensity MAX_I (typically 255 for 8-bit images) and the MSE:

$$\mathrm{PSNR} = 10 \cdot \log_{10} \left(\frac{MAX_I^2}{\mathrm{MSE}} \right) . \tag{2.18}$$

The **Structural Similarity Index Measure (SSIM)** evaluates perceptual image quality by combining luminance, contrast, and structural comparisons over local windows. Given means μ_Y, μ_T,

standard deviations σ_Y, σ_T, and cross-covariance σ_{YT}, SSIM is defined as:

$$\text{SSIM}(Y,T) = \frac{(2\mu_Y\mu_T + c_1)(2\sigma_{YT} + c_2)}{(\mu_Y^2 + \mu_T^2 + c_1)(\sigma_Y^2 + \sigma_T^2 + c_2)}, \tag{2.19}$$

where c_1 and c_2 are constants that stabilize the division and are typically defined as $c_1 = (0.01 \cdot MAX_I)^2$ and $c_2 = (0.03 \cdot MAX_I)^2$.

Together, these three metrics provide a comprehensive assessment of pixel-level accuracy (MSE), perceptual similarity (SSIM), and global image fidelity (PSNR), offering a robust quantitative framework for evaluating enhancement quality in endoscopic imaging.

2.4.2 Experimental Protocol ◂

Experiments that include comparative assessments of traditional exposure correction methods, convolutional neural networks, and models based on transformers were conducted. The results were thoroughly evaluated within a training and testing framework using the Endo-4IE dataset.

Four *traditional methods*—namely RLBHE, FHSABP, LIME, and DUAL—were executed in a non-learning configuration and applied directly to the test images. These baselines serve to establish a lower-bound reference for enhancement quality and computational cost.

The *Endo-LMSPEC model* was trained using a two-phase *patch-wise strategy* to balance memory efficiency and photometric adaptation. Initially, the network was trained on 128×128 non-overlapping pixel patches for 40 epochs, enabling fast convergence on local illumination corrections. This was followed by a fine-tuning stage using 256×256 pixel patches over 30–50 additional epochs, depending on exposure type, to incorporate broader spatial context. Patch-based training was chosen over full-frame learning due to GPU memory constraints and its inherent regularization effect. Overlapping patches (from 25 up to 50% of their surface) were used to reduce boundary artifacts and mitigate tiling effects during inference, particularly at higher frequency bands. The architecture processes Laplacian pyramid decompositions of the input, assigning each frequency level to a dedicated U-Net-based sub-network. This enables coarse-to-fine correction: lower levels manage global exposure, while higher levels refine textures. Adversarial supervision was selectively applied at intermediate scales to enhance realism without compromising structural integrity. Training was supervised using the composite loss function defined in Eq. (2.7), which integrates four components: pixel-wise reconstruction ($\mathscr{L}_{\text{rec}}$), perceptual pyramid loss ($\mathscr{L}_{\text{pyr}}$), structural similarity loss ($\mathscr{L}_{\text{SSIM}}$), and adversarial loss ($\mathscr{L}_{\text{adv}}$). The regularization weights were empirically set as $\alpha = \beta = \delta = 0.25$ and $\gamma = 1.0$ to prioritize perceptual and structural fidelity while preserving adversarial realism. Three specialized variants of the Endo-LMSPEC model were trained to handle different exposure scenarios: underexposed, overexposed, and mixed (i.e., containing both under- and overexposed frames). This exposure-specific tuning ensures that the model adapts optimally to

the photometric distribution present in each subset of the Endo4IE dataset.

The *Endo-ViT model* was trained on the complete frames (resolution of 512×512 pixels) using an Adam optimizer, cosine learning rate decay, and horizontal/vertical data augmentation. The model was trained for up to 50 epochs, with optimal performance observed at the final epoch. Its architecture consists of a dual-branch transformer backbone: a local branch predicts spatial exposure corrections using ViT and PEM modules, while a global branch emulates camera-level ISP operations. The model was supervised using a compound loss function that integrates pixel-wise regression, Laplacian pyramid loss, and a histogram-aware color loss adapted from HistoGAN, with regularization weights $\lambda = 0.1$ and $\beta = 0.04$ in Eq. (2.13).

All models were trained and evaluated using the same data split: 70% for training, 27% for testing, and 3% for validation. In addition to the quantitative evaluation based on PSNR and SSIM (see Section 2.4.1), qualitative comparisons were performed to assess the preservation of anatomical structures, continuity of fine details, and perceptual color accuracy. The evaluation protocol also included ablation studies to isolate and quantify the individual contributions of key architectural components and loss terms within each model.

2.4.3 Model Training and Hyper-Parameter Optimization ◂

Ablation studies were systematically conducted for both DL-models (Endo-LMSPEC and Endo-ViT) to evaluate the impact of individual loss components and architectural features on performance. Each model variant was trained on the same subset of the Endo-4IE dataset and assessed under identical conditions on the corresponding test split, ensuring fair and reproducible comparisons.

For the Endo-LMSPEC model, an initial ablation study focused on the relative contributions of the four loss terms defined in Equation (2.7). Through empirical analysis, the optimal configuration was found to be $\alpha = \beta = \delta = 0.25$ and $\gamma = 1.0$, highlighting the importance of the SSIM term in preserving structural details across varying illumination conditions. Following this, three distinct model variants were trained to specialize in correcting (i) underexposed (UE), (ii) overexposed (OE), and (iii) mixed-exposure (C) frames. Each variant was trained using a two-phase patch-wise approach. In the first phase, the network was trained with 128×128 pixel patches for 30 to 50 epochs to capture local photometric patterns. The second phase fine-tuned the model on 256×256 pixel patches, initialized with weights from the first phase, to incorporate broader contextual information. Patches were extracted from the images with an overlap from 25 up to 50% of their surfaces during training and inference to mitigate edge inconsistencies. Adversarial supervision was introduced in the second phase, with the discriminator activated at a specific discriminator starting epoch (DSE), transitioning the network into a GAN-like configuration. Full hyper-parameter settings, including batch size, learning rates, and DSE values, are detailed in Table 2.1.

The Endo-ViT model followed a similar tuning methodology. Loss component ablations were

TABLE 2.1 – Hyper-parameter configurations for Endo-LMSPEC. Phase 1 uses 128×128 pixel patches; Phase 2 uses 256×256 pixel patches.

Model	Training Set	Phase	Epochs	DSE	BS	lr_G	lr_D
LMSPEC (Afifi et al. [2021])	UE, OE, C	1	40	–	32	10^{-4}	10^{-5}
Endo-LMSPEC	UE, OE, C	2	30	15	8	10^{-4}	10^{-5}
Best (UE)	UE	1	50	–	32	10^{-4}	10^{-5}
		2	40	20	8	10^{-4}	10^{-5}
Best (OE)	OE	1	40	–	64	2×10^{-4}	2×10^{-5}
		2	30	15	32	2×10^{-4}	2×10^{-5}
Best (C)	Combined (UE+OE)	1	50	–	32	10^{-4}	10^{-5}
		2	40	20	8	10^{-4}	10^{-5}

Notes: DSE = discriminator starting epoch. BS = batch size. lr_G/lr_D = learning rate for generator/discriminator.

used to assess the individual contributions of the pixel-wise $\mathscr{L}_1$ loss, Laplacian pyramid loss, and histogram color loss (see Eq. (2.13)). The final configuration adopted fixed weights of $\lambda = 0.1$ and $\beta = 0.04$, which offered the best trade-off between detail preservation and color consistency. Given its lightweight architecture, Endo-ViT required fewer epochs and was trained on full images instead of patches, enabling real-time performance while maintaining accuracy.

Together, these studies guided the architectural and objective function design of both models, enabling a rigorous and performance-driven optimization process.

2.4.3.1 Endo-LMSPEC ◄

Table 2.2 summarizes the results of an ablation study conducted to assess the individual contribution of each loss term included in the full optimization objective defined in Eq. (2.7). Removing the structural similarity loss ($\mathscr{L}_{\text{SSIM}}$) led to a substantial drop in SSIM scores, underscoring its critical role in preserving fine anatomical detail and edge continuity—both essential for maintaining structural integrity in endoscopic images. Similarly, excluding the adversarial loss term ($\mathscr{L}_{\text{adv}}$) degraded the visual realism of the reconstructions, resulting in blurrier textures and diminished contrast in high-frequency regions. This degradation was evident in both PSNR and SSIM, confirming that adversarial supervision contributes to the generation of photorealistic and perceptually coherent outputs.

2.4.3.2 Endo-ViT ◄

Table 2.3 presents results for the ablation of key loss terms and architectural components in the Endo-ViT model. The removal of the histogram loss led to a significant color shift and degraded perceptual quality, as reflected in SSIM scores. The Laplacian pyramid loss was also critical in preserving sharpness and suppressing blurring. Additionally, a variant without the global branch (ISP emulation) showed performance degradation in frames with uneven lighting, emphasizing the complementary nature of the dual-branch architecture.

TABLE 2.2 – Ablation study on Endo-LMSPEC loss components (overexposed subset).

Configuration	**PSNR ↑**	**SSIM ↑**	**MSE ↓**
Full Loss ($\mathscr{L}_{\text{total}}$)	**22.70**	**0.801**	**0.046**
w/o $\mathscr{L}_{\text{SSIM}}$	21.85	0.773	0.053
w/o $\mathscr{L}_{\text{adv}}$	22.01	0.785	0.050
w/o $\mathscr{L}_{\text{pyr}}$	21.69	0.777	0.055

TABLE 2.3 – Ablation study on Endo-ViT components and supervision (underexposed subset).

Configuration	**PSNR ↑**	**SSIM ↑**	**MSE ↓**
Full Model	**24.69**	**0.813**	**0.038**
w/o $\mathscr{L}_{\text{hist}}$	23.46	0.774	0.043
w/o $\mathscr{L}_{\text{pyr}}$	23.12	0.759	0.046
w/o Global Branch	22.80	0.748	0.048
w/o Local Branch	22.26	0.730	0.051

The ablation results demonstrate that both structural and perceptual components of the loss functions play a crucial role in achieving optimal performance. In Endo-LMSPEC, the SSIM loss is particularly effective in preserving anatomical detail and maintaining structural coherence across regions of varying exposure. Adversarial supervision further enhances local contrast, contributing to more realistic texture reconstruction. For Endo-ViT, the histogram-aware loss promotes color fidelity by aligning the chromatic distribution of the enhanced images with the ground truth, while the Laplacian pyramid loss reinforces spatial sharpness by preserving fine-scale details. Additionally, the model's dual-branch architecture—combining local transformer blocks with a global ISP-inspired pathway—enables the simultaneous modeling of fine-grained and global illumination effects, which is especially advantageous under complex, heterogeneous lighting conditions commonly encountered in endoscopy.

2.4.4 Quantitative Evaluation ◂

A quantitative evaluation was performed on the test partition of the Endo-4IE dataset using the standard reference-based image quality metrics described in Section 2.2.1. Results are reported separately for overexposed and underexposed subsets in Table 2.4.

Among the traditional methods, histogram equalization approaches (RLBHE and FHSABP) achieve higher SSIM scores compared to Retinex-based techniques such as LIME and DUAL. However, their PSNR values remain relatively low, indicating limited fidelity in absolute intensity reconstruction. These methods are generally effective in preserving edge information but often fail to correct global tonal inconsistencies, particularly in frames containing mixed illumination zones.

In contrast, the proposed deep learning-based models, Endo-LMSPEC and Endo-ViT, systema-

TABLE 2.4 – Quantitative results on the Endo-4IE test set. Metrics are reported separately for overexposed and underexposed frames. Best results per column are highlighted in bold.

Method	PSNR ↑	SSIM ↑	MSE ↓	Inference Time (s)
Overexposed frames				
RLBHE	19.43	0.746	0.051	0.2996
FHSABP	16.02	0.631	0.036	0.2953
LIME	8.57	0.597	0.048	44.0236
DUAL	0.91	0.726	0.043	30.9965
Endo-LMSPEC	22.70	0.801	0.046	0.1316
Endo-ViT	**23.23**	**0.824**	**0.038**	**0.0135**
Underexposed frames				
RLBHE	21.05	0.723	0.055	0.2996
FHSABP	18.20	0.633	0.034	0.2953
LIME	17.33	0.699	0.054	44.0236
DUAL	20.01	0.708	0.053	30.9965
Endo-LMSPEC	23.23	0.786	0.046	0.1316
Endo-ViT	**24.69**	**0.813**	**0.038**	**0.0135**

tically outperform classical methods across all metrics. Endo-ViT achieves the highest PSNR and SSIM scores for both exposure conditions, while maintaining a real-time inference speed of 74 FPS (corresponding to 0.0135 seconds per frame). Notably, Endo-ViT reduces the mean squared error (MSE) substantially, indicating high pixel-wise reconstruction accuracy. These results confirm the superiority of transformer-based architectures for exposure correction in endoscopic imaging, particularly in images including complex illumination variations.

2.4.5 Qualitative Evaluation ◂

In addition to the quantitative evaluation, a comprehensive qualitative analysis was conducted to assess the perceptual and anatomical quality of the enhanced images under diverse photometric conditions. This analysis focuses on three key dimensions: (i) accurate restoration of exposure levels, which directly affects perceived contrast and visibility; (ii) preservation of structural detail, particularly at anatomical boundaries; and (iii) chromatic consistency, ensuring that color reproduction remains realistic and diagnostically coherent across different enhancement methods.

Comparison with Traditional Methods. Figure 2.11 presents a comparative analysis between traditional image enhancement techniques and the proposed Endo-LMSPEC model across various exposure conditions. Classical methods—such as histogram equalization and Retinex-based approaches—offer limited correction and frequently lead to over-saturation or spatially uneven results, especially in regions affected by non-uniform illumination. In contrast, Endo-LMSPEC effectively

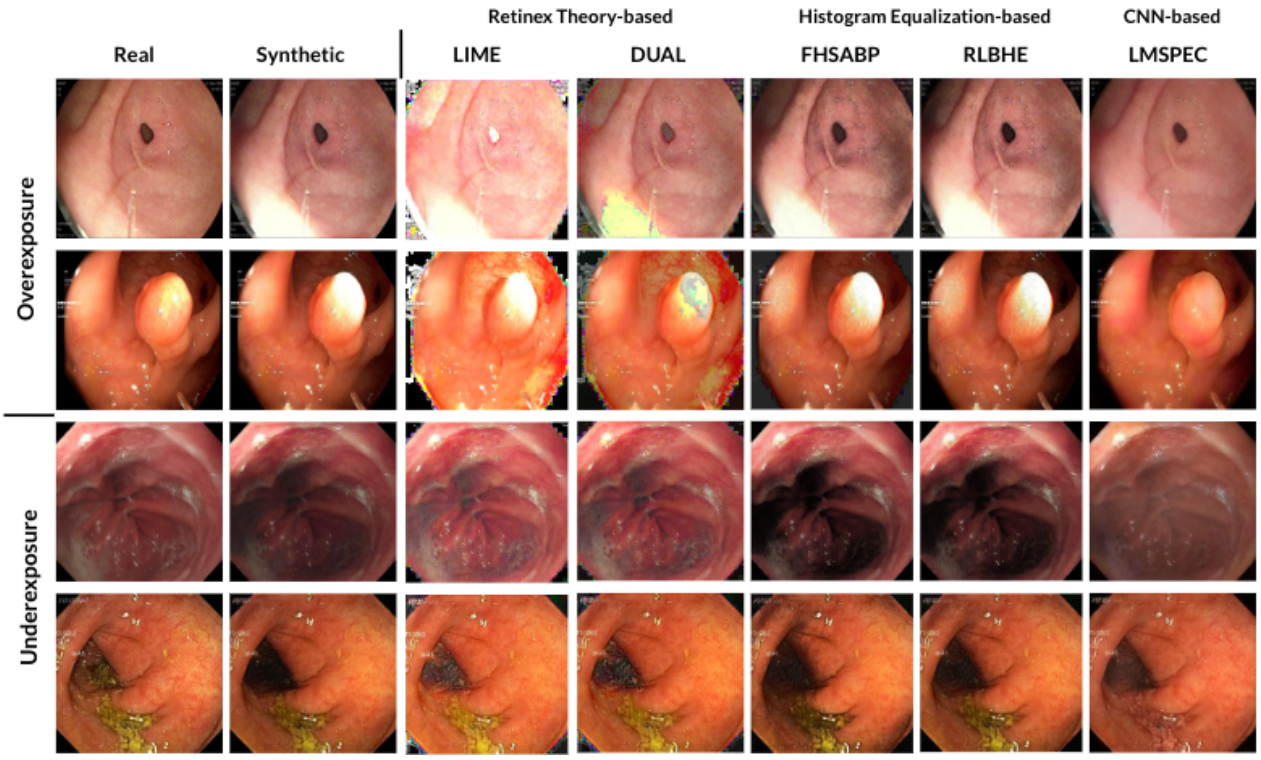


FIGURE 2.11 – Comparison between traditional enhancement methods and Endo-LMSPEC on the Endo-4IE dataset. From left to right: acquired reference image, synthetically exposure corrupted images, and enhanced outputs from Retinex-based, histogram-based, and CNN-based (Endo-LMSPEC) methods. Endo-LMSPEC better preserves structures and the enhanced under- or over-exposed regions are more artifact-free.

restores anatomical structures and preserves the overall visual coherence of the image across both overexposed and underexposed frames. Its ability to adapt to local brightness variations without introducing artifacts enhances both perceptual quality and anatomical clarity.

Comparison between LMSPEC and Endo-LMSPEC. Figure 2.12 provides a qualitative comparison between the original LMSPEC method and the proposed Endo-LMSPEC architecture, evaluated under two representative photometric degradation scenarios: overexposure (top row) and underexposure (bottom row). The figure presents, in sequence, the non-degraded reference frame, the corrupted input, the LMSPEC-enhanced result, and the output generated by Endo-LMSPEC.

Although LMSPEC is capable of attenuating gross exposure-related artifacts, its corrections are often suboptimal in terms of contrast enhancement and color fidelity—particularly in overexposed areas where high-intensity regions tend to remain saturated and washed out. In comparison, Endo-LMSPEC demonstrates superior recovery of global contrast and finer anatomical features, leading to outputs that are both visually coherent and clinically meaningful. These improvements are especially evident in the zoomed-in regions, where Endo-LMSPEC exhibits enhanced texture preservation and suppresses residual artifacts that remain visible in the LMSPEC outputs.

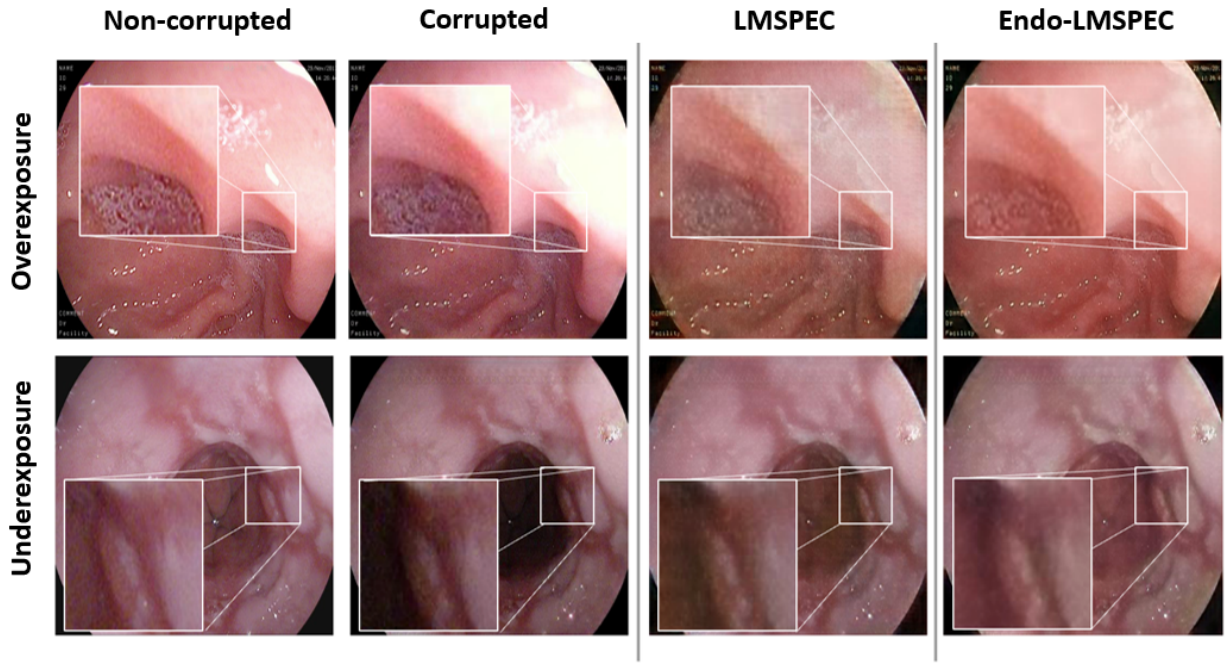


FIGURE 2.12 – Enhancement results from Endo-LMSPEC for two photometric degradation scenarios: overexposure (top) and underexposure (bottom). The model effectively restores contrast, anatomical structures, and boundary contours. However, minor hue shifts may persist in regions severely affected by exposure extremes, indicating residual color correction challenges in highly saturated or dim areas.

Comparison between Endo-LMSPEC and Endo-ViT. Figure 2.13 allows for a visual comparison of the results obtained with Endo-LMSPEC and the transformer-based Endo-ViT model. Both models improve exposure uniformity and anatomical clarity, yet Endo-ViT consistently produces sharper boundaries and more perceptually natural textures. For example, in the underexposed region (top row), Endo-ViT more accurately restores the contrast between tissue boundaries and background mucosa without amplifying specular reflections. In the overexposed case (bottom row), the transformer-based model reduces residual shadows and restores subtle glandular textures, which appear smoothed or muted in the Endo-LMSPEC output. These differences can be attributed to Endo-ViT's self-attention mechanism and dual-branch design, which jointly encode both local and global illumination cues.

Histogram-Based Color Consistency. Figures 2.14 and 2.15 compare color histograms respectively for overexposed and underexposed frames to assess the chromatic fidelity power of the proposed DL approaches. In both scenarios, the histograms of the images generated by Endo-ViT exhibit the highest color distribution similarity with the histograms of the ground truth, particularly in the red and green channels. This effect is especially visible in specular regions, where Endo-LMSPEC tends to introduce minor chromatic drift or desaturation. These findings empirically validate the

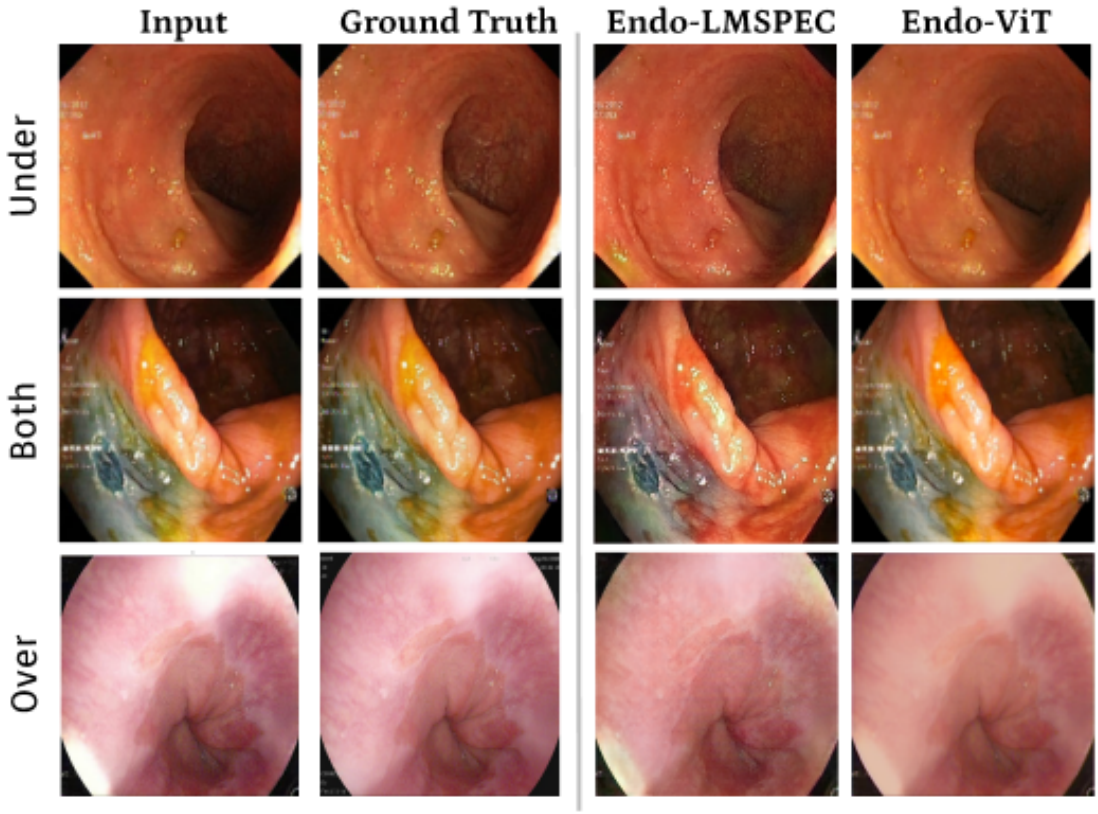


FIGURE 2.13 – Visual comparison between Endo-LMSPEC and Endo-ViT for overexposed (bottom), underexposed (top), and mixed exposure frames (middle). From left to right: acquired image; ground truth (exposure-corrected); Endo-LMSPEC output; Endo-ViT output. Endo-ViT better restores tissue contrast and anatomical boundaries across all cases, especially in high-frequency regions with strong illumination gradients or specular artifacts.

contribution of the histogram-aware loss in Endo-ViT, which enforces global chromatic alignment during training.

Concluding remarks on the qualitative assessment. Across all visual comparisons, Endo-ViT consistently demonstrates improved exposure correction, structural sharpness, and chromatic consistency relative to both classical methods and Endo-LMSPEC. These qualitative results reinforce the findings from the quantitative evaluation, highlighting the effectiveness of transformer-based architectures and histogram-aware training objectives in addressing complex illumination artifacts in endoscopic imaging.

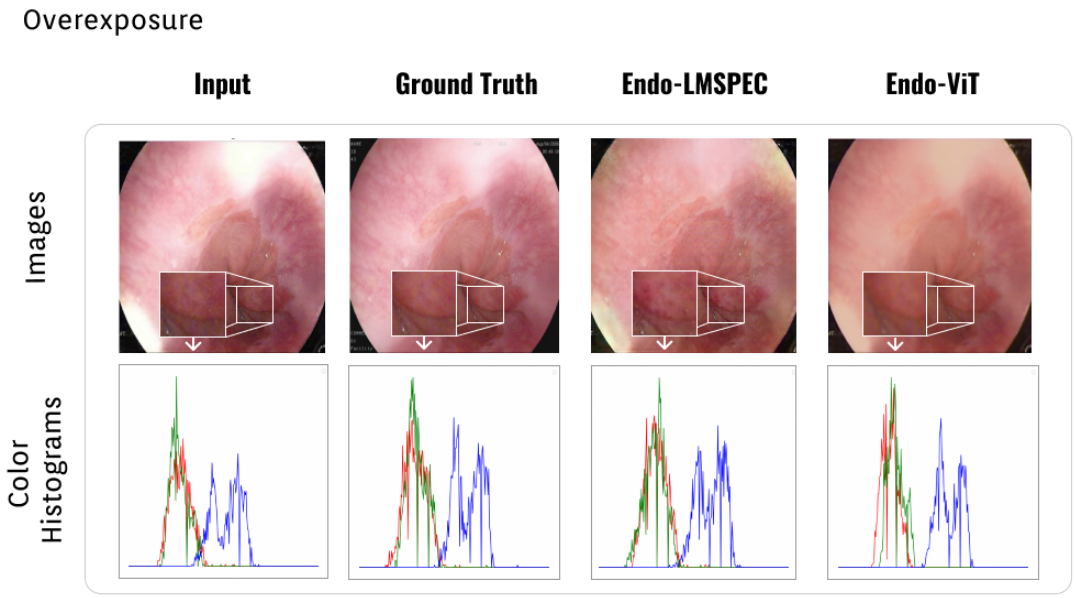


FIGURE 2.14 – Qualitative comparison of enhancement results in overexposed regions using RGB histograms. Top row (left to right): input image, ground truth reference, image corrected by Endo-LMSPEC, and image corrected by Endo-ViT. Bottom row: corresponding RGB histograms computed from the zoomed-in region indicated above. Endo-ViT demonstrates closer chromatic alignment with the ground truth, particularly in the red and green channels.

2.5 Conclusion ◂

This chapter addressed the challenge of correcting illumination artifacts in endoscopic imaging, with a focus on overexposed and underexposed image regions commonly encountered in gastrointestinal procedures. These photometric degradations were shown to negatively affect both visual interpretability and the reliability of computer-assisted analysis pipelines.

The Endo-4IE dataset was developed using real endoscopic frames and synthetically generated exposure errors through GAN-based translation to enable the supervised training of enhancement models. The resulting paired dataset allows for structured evaluation under a variety of exposure conditions and serves as a robust benchmark for both traditional and learning-based correction techniques.

Two deep learning-based enhancement models were evaluated in detail: Endo-LMSPEC, a convolutional multi-scale architecture based on Laplacian pyramid decomposition, and Endo-ViT (Espinosa et al. [2024b]), a transformer-based model that introduces color-aware supervision and a dual-branch correction strategy. Quantitative results demonstrated that Endo-ViT systematically outperforms baselines across multiple metrics, including PSNR and SSIM.

Ablation studies confirmed the contribution of key components in both architectures. In Endo-LMSPEC, the SSIM term played a critical role in preserving structural integrity, while adversarial losses promoted perceptual realism. In Endo-ViT, the histogram-aware loss and Laplacian pyra-

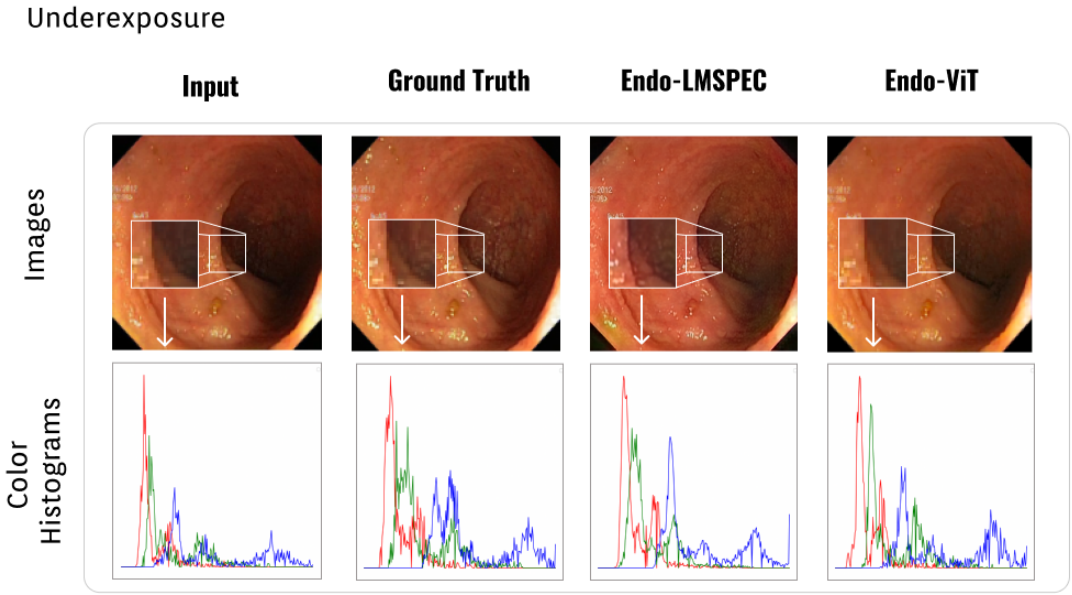


FIGURE 2.15 – Qualitative comparison for underexposed regions using histograms. Top row from the left to the right: input image, ground truth, Endo-LMSPEC corrected image, Endo-ViT corrected. Bottom: RGB histograms of the zoomed region (boxed above). Endo-ViT restores low-light detail and preserves color distributions more faithfully than Endo-LMSPEC.

mid supervision proved essential for maintaining both color fidelity and edge sharpness. Qualitative evaluation further revealed that Endo-ViT produced enhanced images with improved visual consistency, particularly in color-sensitive regions, as confirmed by histogram alignment with reference distributions.

In addition to improved accuracy, Endo-ViT demonstrated superior computational efficiency, requiring fewer than 100K parameters and supporting real-time inference at over 70 frames per second. This performance highlights the practical suitability of transformer-based exposure correction for integration into real-time clinical systems.

The next chapter examines how enhanced frames can be exploited to improve geometric consistency and reconstruction quality in 3D modeling and navigation pipelines, such as SLAM and Structure-from-Motion.

Publications related to chapter 2.

— **A Novel Hybrid Endoscopic Dataset for Evaluating Machine Learning-Based Photometric Image Enhancement Models**
Carlos Axel Garcia-Vega, Ricardo Espinosa, Gilberto Ochoa-Ruiz, Thomas Bazin, Luis Eduardo Falcón-Morales, Dominique Lamarque, Christian Daul
Advances in Computational Intelligence – MICAI 2022, Lecture Notes in Computer Science, vol. 13612, Springer, 2022, pp. 267–281.

— **Multi-Scale Structural-Aware Exposure Correction for Endoscopic Imaging**
A. Garcia-Vega, R. Espinosa, L. Ramirez-Guzman, T. Bazin, L. Falcón-Morales, G. Ochoa-Ruiz, D. Lamarque, and C. Daul.
In: *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, 2023.

— **Color-Aware Exposure Correction for Endoscopic Imaging Using a Lightweight Vision Transformer**
Ricardo Espinosa, Javier Eluney Hernández, Gilberto Ochoa-Ruiz, Christian Daul
IEEE 37th International Symposium on Computer-Based Medical Systems (CBMS), 2024, pp. 376–382.

— **Prompt Assisted Enhancement for Correcting Illumination Artifacts in Endoscopic Images**
Ricardo Espinosa, Gilberto Ochoa-Ruiz, Christian Daul
Accepted for oral presentation at MICAI 2025. To be published in the Springer LNAI series.

Chapter 3

Integration of DL-based Image Enhancement in 3D Reconstruction Pipelines for Colonoscopy

---

3.1 Introduction ◂

Deep learning-based 3D reconstruction approaches have demonstrated significant potential in augmenting endoscopic imaging capabilities, particularly by expanding the observable regions of internal organ surfaces. Despite these advances, robust and accurate 3D reconstruction from endoscopic video sequences remains a challenging task. This difficulty arises primarily from adverse photometric conditions—such as localized overexposure, specular highlights, and rapid illumination changes induced by the dynamic light source positioned at the tip of the endoscope—compounded by the highly reflective mucosal surfaces within the gastrointestinal tract. These photometric artifacts frequently violate the foundational assumptions of conventional 3D reconstruction pipelines, which typically rely on classical Structure-from-Motion (SfM) or Simultaneous Localization and Mapping (SLAM) frameworks. Consequently, these violations impair camera pose estimation accuracy and lead to inconsistent and unreliable surface geometry reconstructions.

To mitigate the aforementioned limitations, this chapter explores the integration of deep learning-based image enhancement techniques as a dedicated preprocessing stage within monocular 3D reconstruction pipelines. Specifically, we incorporate the DL-based enhancement modules introduced in Chapter 2, namely, Endo-LMSPEC and Endo-STTN to address photometric inconsistencies and improve the depth estimation under challenging illumination conditions. In this framework, Endo-LMSPEC is employed to locally normalize exposure across frames, thereby promoting structural coherence throughout the endoscopic video sequence. Simultaneously, Endo-STTN is utilized to inpaint specular reflections, effectively restoring surface information occluded by these artifacts and enabling more reliable geometry inference.

This chapter presents the integration of a deep learning-based enhancement method into a SLAM pipeline for colonoscopic surface reconstruction as first introduced in (Espinosa et al. [2023]), where DL-based exposure correction was proposed as a pre-processing stage for improving monocular 3D reconstruction in endoscopy.

The proposed methodology builds upon the RNN-SLAM architecture (Ma et al. [2021]), a deep learning-based framework specifically developed for monocular 3D reconstruction in endoscopic imaging. This architecture integrates recurrent neural networks within a SLAM-inspired pipeline to jointly estimate camera motion and reconstruct dense surface geometry from monocular video sequences. Its design is particularly well-suited for colonoscopic data, as it effectively captures

temporal correlations across frames and leverages data-driven priors to address challenges such as texture sparsity and non-rigid tissue deformation inherent to the gastrointestinal tract. In the present work, the RNN-SLAM pipeline is extended by incorporating the previously introduced deep learning-based image enhancement modules as a preprocessing stage. This integration aims to mitigate photometric inconsistencies and improve the spatial coherence of the reconstructed 3D surfaces.

The chapter is organized as follows: Section 3.2 outlines the SLAM-based 3D reconstruction method chosen for this thesis. Section 3.3 outlines our approach and explains the operation of DL-based image enhancement techniques, as well as their incorporation into a comprehensive 3D reconstruction pipeline. Lastly, Section 4.5 provides both quantitative and qualitative assessments that illustrate the advantages of this integration on realistic colonoscopic sequences.

3.2 Overview of the RNN-SLAM Pipeline ◂

The RNN-SLAM framework (Ma et al. [2021]) combines the strengths of traditional geometric SLAM methods with the temporal modeling capacity of recurrent neural networks to achieve robust monocular depth and pose estimation in unconstrained environments. This hybrid strategy addresses the limitations of both classical SLAM—typically reliant on consistent lighting and distinct textures—and purely learning-based methods that often lack geometric consistency over time.

Figure 3.1 illustrates the main stages of the RNN-SLAM pipeline. The architecture integrates a recurrent depth prediction module (RNN-DP) within a direct sparse odometry (DSO)-inspired SLAM backend. This combination enables the pipeline to simultaneously infer scene geometry and camera motion while leveraging sequential dependencies.

Recurrent Depth Prediction (RNN-DP). At each incoming time step, the input frame is passed through the RNN-DP module, which produces a dense depth prediction conditioned not only on the current image but also on the hidden state accumulated from previous frames. This hidden

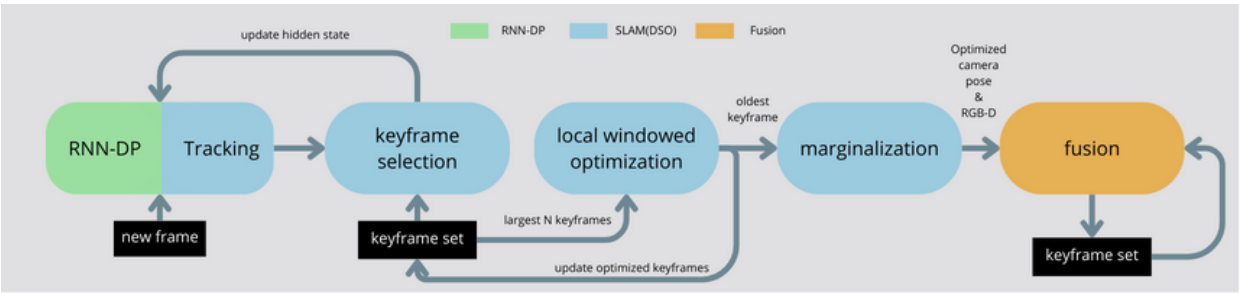


FIGURE 3.1 – Overview of the RNN-SLAM pipeline integrating a recurrent depth prediction module (RNN-DP) with classical SLAM components such as tracking, local optimization, and keyframe-based fusion. Adapted from Ma et al. (Espinosa et al. [2024a])

state encodes temporal information, allowing the network to infer depth based on both motion and contextual continuity over time. The depth prediction serves as the initial geometry prior for the subsequent pose estimation block.

Tracking and Keyframe Selection. The predicted depth and current image are used by the tracking module to estimate the relative camera pose with respect to the most recent keyframe. The pose estimation module aims to minimize the photometric error between the warped keyframe and the current frame by integrating geometric and appearance cues. Subsequently, the current frame can be selected as a keyframe and added to the active set, based on estimates of the motion magnitude and photometric diversity.

Local Windowed Optimization. Once a new keyframe is introduced, the system performs a local bundle adjustment over a temporal window, consisting of the most recent N keyframes. This optimization jointly refines both camera poses and depth maps by minimizing multi-view photometric residuals. Such optimization increases the temporal consistency and geometric accuracy while maintaining computational tractability.

Marginalization and Global Fusion. To limit computational burden, the oldest keyframe in the optimization window is marginalized after each iteration. The optimized RGB-D keyframes are then integrated into a global model using a fusion mechanism that accumulates both color and depth information. This process allows the generation of dense 3D reconstructions over extended image sequences.

Advantages and Limitations. The RNN-SLAM pipeline is particularly well-suited to dynamic and complex visual environments due to its hybrid design. The use of an RNN improves temporal coherence in depth predictions, while local optimization ensures geometric stability. However, the framework is highly sensitive to artifacts in the input images. Photometric inconsistencies—such as specularities, shadows, and variable exposure—can degrade the tracking performance and affect the reliability of the optimization and fusion stages.

To address these limitations, this contribution incorporates photometric pre-processing modules, namely Endo-STTN and Endo-LMSPEC, as a front-end enhancement step. These modules correct exposure variations and enhance local contrast, thereby improving the robustness of RNN-SLAM under challenging illumination conditions typical in colonoscopy. The impact of this enhancement on tracking stability, depth consistency, and surface reconstruction quality is discussed in the subsequent sections.

3.3 Proposed 3D Reconstruction Pipeline ◂

The proposed 3D reconstruction pipeline, depicted in Figure 3.2, extends the RNN-SLAM framework by integrating deep learning-based photometric enhancement modules as a dedicated pre-processing stage. While the core architecture—comprising visual odometry, recurrent pose tracking, and volumetric fusion—has been detailed in the preceding section, this extension specifically targets the mitigation of photometric distortions commonly encountered in colonoscopic imagery. By addressing illumination inconsistencies and artifact-induced signal degradation prior to keyframe selection, depth estimation, and fusion, the overall fidelity and geometric consistency of the reconstructed surfaces are significantly improved.

Three deep learning-based photometric enhancement modules were evaluated as pre-processing components: *Endo-LMSPEC* for correcting under- and overexposed regions, *Endo-STTN* for inpainting specular reflections, and a standard *Gamma Correction* module for global contrast adjustment. These components are applied directly to the incoming video frames prior to the pose tracking stage. The illumination normalization module (*Endo-LMSPEC*) is designed to locally compensate for exposure variations throughout the video sequence, thereby improving temporal consistency and preserving structural details. The specularity inpainting module (*Endo-STTN*) addresses high-frequency artifacts caused by reflective mucosal surfaces, effectively recovering the underlying texture information. The *Recurrent Gamma Correction* module performs a non-linear intensity transformation to enhance global contrast across frames, with particular emphasis on improving visibility in underexposed regions. Unlike traditional gamma correction, this module incorporates temporal feedback through a recurrent structure, enabling it to adaptively adjust the gamma response based on the illumination dynamics observed over time. As demonstrated in Espinosa et al. [2024a]; Zhang et al. [2021], this pre-conditioning strategy enhances the robustness of the reconstruction pipeline by ensuring that the visual input remains photometrically stable and geometrically consistent.

This section introduces the image enhancement techniques evaluated in the pipeline and discusses their integration and their effect on 3D reconstruction quality in colonoscopic video se-

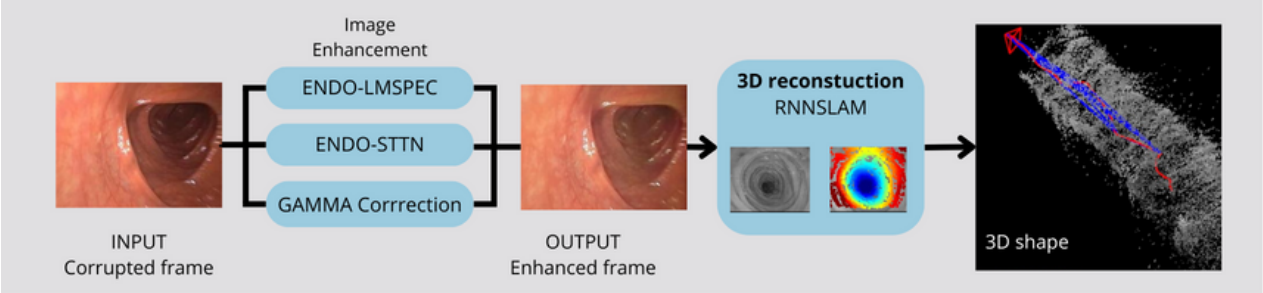


FIGURE 3.2 – Proposed pipeline including Endo-LSMPEC, Endo-STTN, and Recurrent Gamma Correction methods as image enhancement afterward 3D reconstruction using RNN-SLAM module. (Espinosa et al. [2024a])

quences.

3.3.1 Explored Image Enhancement Methods ◄

To improve photometric quality prior to geometric reconstruction, three image enhancement methods are evaluated as pre-processing modules: (i) Recurrent Gamma Correction for global contrast enhancement, (ii) Endo-STTN for specularity inpainting, and (iii) Endo-LMSPEC for localized exposure normalization. Each method addresses different types of photometric distortions commonly encountered in colonoscopic sequences.

(i) **Endo-LMSPEC**, a multi-scale structural-aware exposure correction model proposed in this thesis and (Vega et al. [2023]);

(ii) **Endo-STTN**, a transformer-based temporal inpainting method for specular artifact removal (Daher et al. [2023]); and

(iii) **Recurrent Gamma Correction**, a recurrent neural network approach for video-level illumination normalization (Zhang et al. [2021]).

The deep learning-based methods (i) and (ii) are presented in full detail in Chapter 2, including their architectural design, training procedures, and enhancement performance. This section focuses on how each method contributes to photometric consistency in video sequences and impacts downstream 3D reconstruction accuracy.

3.3.1.1 Multi-Scale Exposure Correction (Endo-LMSPEC) ◄

Endo-LMSPEC (Chapter 2) is a Laplacian pyramid-based convolutional model designed to correct local exposure distortions at multiple spatial frequencies. By decomposing each image into frequency bands and processing them with specialized U-Net-inspired subnetworks, the model restores both global contrast and local structural detail. When applied as a pre-processing stage, it improves photometric uniformity across frames, reducing visual noise that may otherwise hinder depth estimation and SLAM tracking.

3.3.1.2 Recurrent Gamma Correction ◄

The gamma correction method introduced in (Zhang et al. [2021]) employs a recurrent neural network to estimate adaptive, frame-specific gamma values across video sequences. Unlike conventional static gamma correction, this approach leverages temporal dependencies by incorporating contextual information from previous frames to modulate the correction applied to the current frame. The goal is to achieve temporally coherent illumination normalization that accounts for both local intensity distribution and global exposure trends throughout the sequence. The correction is defined as:

$$I_o = A \cdot I_i^{\gamma}, \tag{3.1}$$

where I_i denotes the input image, γ is the gamma parameter dynamically estimated by the network for each frame, and A is a global scaling constant to preserve intensity range. By adjusting the non-linear intensity response based on temporal feedback, the method effectively suppresses frame-to-frame brightness fluctuations and flickering artifacts. This temporal stability is particularly beneficial in visual SLAM and 3D reconstruction tasks, where consistent photometric appearance is crucial for accurate point cloud alignment, depth estimation, and feature tracking. Experimental results reported in (Zhang et al. [2021]) demonstrate that this recurrent formulation leads to improved robustness in both low-light scenarios and in sequences with highly variable illumination.

3.3.1.3 Temporal Specularity Inpainting (Endo-STTN) ◄

Endo-STTN is an adaptation of the Spatial-Temporal Transformer Network (STTN) (Daher et al. [2023]), originally designed for generic video inpainting. The architecture integrates a transformer-based attention mechanism between convolutional encoder-decoder modules to restore occluded or corrupted video regions by leveraging temporal redundancy across frames.

In the context of endoscopy, we repurpose this model for specularity removal—a critical preprocessing step for 3D reconstruction, as specular highlights can disrupt both photometric consistency and geometric inference. Specularities in endoscopic images often appear as high-intensity regions lacking anatomical information, making them challenging for traditional frame-based enhancement techniques.

To train Endo-STTN for this task, we first generate specularity masks using a segmentation model trained on annotations derived from the Dichromatic Reflection Model (DRM). These masks are expanded using morphological dilation to include surrounding dark halos and are used to synthetically occlude regions in training sequences. The masked areas simulate specular artifacts and serve as the target for inpainting.

The inpainting process relies on attention maps computed across the temporal dimension. For each corrupted frame, the model identifies and transfers content from visually and temporally coherent regions in adjacent frames to fill the occluded area. This temporal consistency allows Endo-STTN to preserve spatial context and motion coherence, even in dynamic scenes.

The final model is fine-tuned on endoscopic video data using adversarial and reconstruction losses, enabling it to generalize to real clinical sequences. When applied as a pre-processing step in a 3D reconstruction pipeline, Endo-STTN effectively eliminates specular discontinuities across frames, reducing geometric misalignments and improving the accuracy of both depth estimation and camera pose tracking.

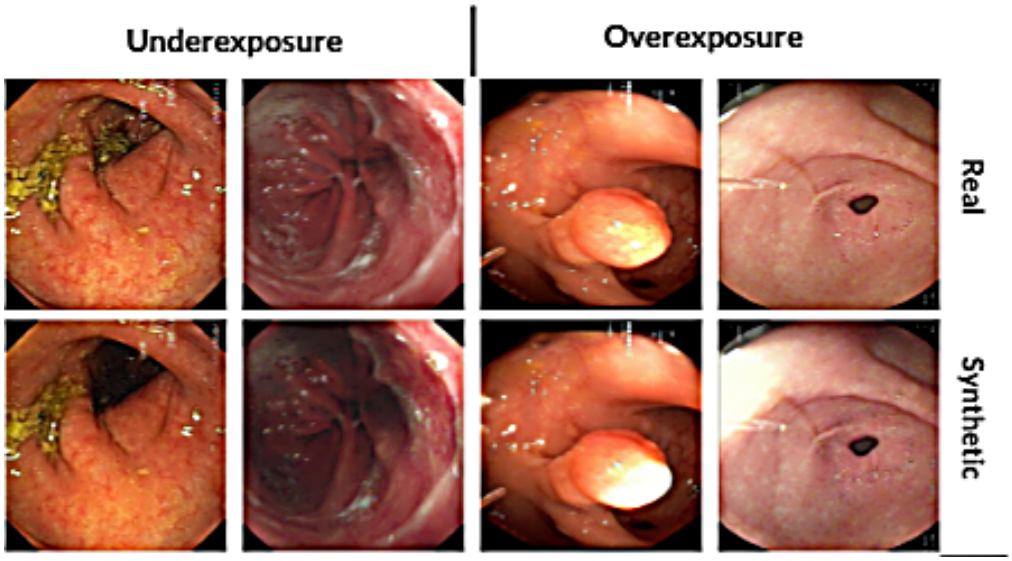


FIGURE 3.3 – Comparison of real and synthetic endoscopic images under varying illumination conditions from Endo4IE dataset. The top row shows real colonoscopic images, while the bottom row presents their synthetic counterparts. Columns are grouped into underexposed and overexposed scenarios, highlighting the diversity and realism of the synthetic dataset in replicating challenging lighting conditions.

3.3.2 Used Datasets ◄

Two datasets were employed to train the different components of the proposed pipeline: (1) the Endo4IE dataset for training the DL-based photometric enhancement models, and (2) a dataset generated using Structure-from-Motion (SfM) for training the recurrent depth prediction network. Additionally, real colonoscopic video sequences were used in the 3D reconstruction phase to evaluate the performance of the pipeline under realistic conditions.

3.3.2.1 Endo4IE Dataset. ◄

To ensure fair training conditions across the three compared image enhancement methods, the Endo4IE dataset proposed in this thesis was employed. An example of this dataset is shown in Figure 3.3. The Endo4IE dataset consists of paired endoscopic frames generated using the CycleGAN architecture for image-to-image translation. This approach enabled the synthesis of overexposed and underexposed versions from unmodified endoscopic frames originally extracted from publicly available datasets, including the EAD2020 (Ali et al. [2020]), EAD2.0 (Ali et al. [2021a]), and HyperKvasir (Borgli et al. [2020]).

The dataset is composed of images, each with a resolution of 512×512 pixels, categorized into three distinct subgroups: (i) 2,216 original images serving as ground truth, (ii) 1,231 images simulated to be overexposed, and (iii) 985 images simulated to be underexposed. Every original image has a matched synthetic counterpart, providing exact pairings for supervised learning. The

dataset was divided into 70% for training, 27% for testing, and 3% for validation.

3.3.2.2 SfM-Generated Dataset ◄

To train the recurrent depth prediction module in the absence of ground truth depth and pose information, a dataset was created using a Structure-from-Motion (SfM) approach, following the procedure described in (Ma et al. [2021]). This technique was applied to reconstruct 3D geometry and estimate camera trajectories from 60 real colonoscopy video sequences, each containing approximately 20,000 frames.

Using a sliding window of 200 consecutive frames, the SfM algorithm estimated corresponding depth maps and camera poses, which were subsequently used as training data for the RNN-based depth prediction network. The generated dataset provides geometrically consistent pseudo-ground truth data, enabling the training of monocular depth estimation models under realistic colonoscopic conditions.

3.3.3 Training and implementation details ◄

Figure 3.2 illustrates how RNN-SLAM utilizes forecasts of the depth and camera position as an initial step for further calculations. The RNN network consists of two parts: a depth estimation network and a camera pose estimation network. The depth estimation network produces a depth map of the same size as the input image, while the camera pose estimation network generates a relative 6-DoF (degree of freedom) camera pose between the current and prior frames. If the camera intrinsic parameters are known, the dense flow field for 2D pixels can be calculated from the current view to the prior view by utilizing the estimated depth map, camera pose, and camera intrinsic parameters. The estimated depth maps and camera poses are then used to generate dense flow fields to warp previous views to the current view through a differentiable geometric module. The training phase was re-implemented on Tensorflow 2.11, using the training set described in section 4.2, in a fully supervised manner over 20 epochs, using a 0.0002 learning rate and Adam as optimizer. The model was fed with 10 frames which are grouped in a sliding window fashion, in order to preserve the temporal information between frames (Wang et al. [2019]).

3.4 Experimental Results ◄

3.4.1 Qualitative Comparison of Enhancement Methods ◄

To complement the quantitative analysis, Figure 3.4 presents a visual comparison of different image enhancement (IE) methods applied to underexposed and overexposed colonoscopic frames.

The figure displays results from four configurations: the original RGB frame, gamma correction, Endo-STTN, and Endo-LMSPEC.

Underexposed Frames. In the top row, the original frame suffers from poor visibility and reduced contrast due to insufficient illumination. The gamma correction approach enhances global brightness but fails to recover detailed structures in the darker regions. Endo-STTN produces a more perceptually coherent output by enhancing fine vascular patterns and reducing shadow artifacts, although some texture flattening is noticeable. Endo-LMSPEC yields the most balanced enhancement by preserving anatomical detail while achieving consistent brightness across the frame. The multi-scale frequency-aware design of Endo-LMSPEC proves effective in correcting localized intensity deficits without amplifying noise or over-brightening low-intensity areas.

Overexposed Frames. The second row highlights the behavior of the methods under overexposure. In the original RGB input, highlight saturation and loss of structural contours are clearly observed. Gamma correction tends to dim the entire image uniformly, resulting in unnatural tone compression. Endo-STTN is able to mitigate specular regions and restore continuity in high-intensity areas, though it occasionally introduces subtle blurring. Endo-LMSPEC again outperforms the other methods, restoring tonal transitions and anatomical contours with minimal chromatic distortion. Its structural guidance and pyramid-based decomposition allow localized adjustments without affecting well-exposed regions.

Overall, Figure 3.4 illustrates the complementary nature of the evaluated methods. While gamma correction is computationally simple, it lacks semantic and structural awareness. Endo-STTN is specialized for temporal consistency and specularity removal, whereas Endo-LMSPEC provides fine-grained correction of exposure anomalies. These observations validate the quantitative findings and support the inclusion of enhancement models as effective preprocessing modules for robust 3D reconstruction.

3.4.2 Trajectory accuracy ◂

The precision of the projected trajectories of the endoscope camera, directly affecting the fidelity of the 3D surface reconstruction, was evaluated with the EVO toolbox (Sharafutdinov et al. [2023]), a popular open-source SLAM assessment tool. To determine how photometric pre-processing and specularity elimination impact the RNN-SLAM pipeline's efficacy, we compared the calculated trajectories to reference ground-truth trajectories. This analysis encompassed both the original pipeline without image modifications and the versions incorporating improved frames.

The reference ground-truth trajectories were generated offline using the COLMAP framework (Fu et al. [2023]), a state-of-the-art Structure-from-Motion (SfM) method based on exhaustive pairwise

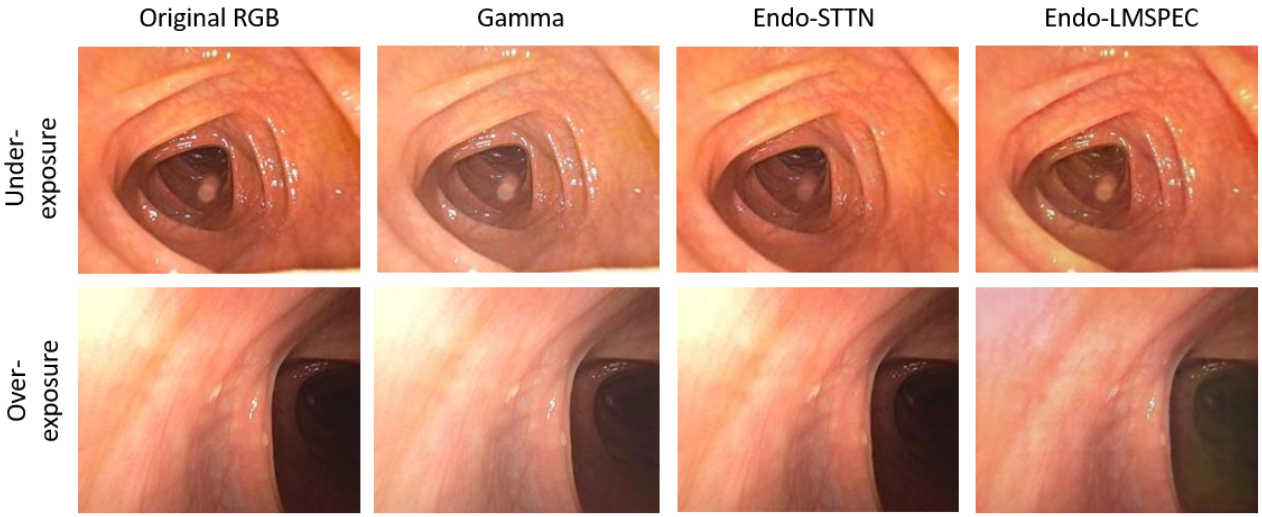


FIGURE 3.4 – Qualitative comparison of image enhancement (IE) methods applied to frames exhibiting underexposure (top row) and overexposure (bottom row). From left to right: original RGB input, Recurrent Gamma Correction, Endo-STTN for specularity inpainting, and Endo-LMSPEC for localized exposure normalization. Endo-LMSPEC shows superior performance in recovering structural details in both low- and high-intensity regions, while Gamma Correction improves global contrast and Endo-STTN reduces the impact of specular highlights.

image matching and global bundle adjustment. Although computationally intensive, COLMAP produces highly accurate camera trajectories, which makes its outputs suitable as a proxy for ground truth in trajectory evaluation. The testing set consisted of four real colonoscopic sequences, selected for their diversity in camera motion and illumination conditions.

To quantify trajectory accuracy, two standard metrics were employed: the Absolute Pose Error (APE) and the Root Mean Square Error (RMSE). These metrics are defined as follows.

Given a ground truth pose P_i^{gt} and an estimated pose P_i^{est} for frame i, the relative pose error E_i is computed as:

$$E_i = (P_i^{gt})^{-1} P_i^{est} \in SE, \tag{3.2}$$

where SE denotes the special Euclidean group representing rigid-body transformations. The translational component of E_i defines the APE at frame i:

$$APE_i = \|\text{trans}(E_i)\|_2. \tag{3.3}$$

To aggregate the trajectory deviation over the sequence, the RMSE is computed as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} APE_i^2}, \tag{3.4}$$

where N denotes the number of frames.

This evaluation protocol provides a robust framework to examine the effect of photometric enhancement techniques on the spatial accuracy of estimated trajectories. By comparing the APE and RMSE values across different pre-processing configurations, the impact of exposure correction and artifact removal on downstream SLAM performance can be rigorously quantified.

3.4.3 Quantitative Results ◂

In this section we discuss our results, summarized in Table 3.1, which shows the metrics essential for assessing the performance of three distinct approaches: RNN-SLAM without image enhancement, RNN-SLAM with gamma correction Daher et al. [2023], RNN-SLAM with Endo-STTN Zhang et al. [2021], and RNN-SLAM with Endo-LMSPEC as proposed in this thesis. The metric values, where lower values indicate better performance, serve as indicators of the proximity between RNN-SLAM trajectories and their ground truth counterparts (Colmap-trajectories). Notably, the proposed method (RNN-SLAM & Endo-LMSPEC) demonstrates superior performance over both RNN-SLAM without enhancement, with gamma correction and specularity inpainting. This is evidenced by the consistently lower metric values, highlighted in bold, in favor of global exposure correction and specularity inpainting methods. Additionally, the results demonstrate that without applying any image enhancement methods to the RNN-SLAM pipeline yields better results in comparison with RNN-Gamma correction and Endo-STTN methods.

When discussing the robustness of the experiments, 4 different datasets were used to gather data. Additionally, considering that RNN-SLAM, as its name suggests, is a recursive method and that better results are likely to be obtained in executions where the algorithm has been run previously, for each experiment (understood as the combination of a dataset and a method), the algorithm was executed multiple times until consistency between executions could be ensured. This approach helped eliminate potential errors due to the runtime environment.

Figure 3.5 complements these quantitative assessments by providing visual representations of the trajectories for a selected colonoscopic video using the three image enhancement methods compared with the baseline method without any image pre-processing. Remarkably, it is evident that the trajectories generated by the proposed SLAM method with image correction using Endo-LMSPEC (depicted as blue lines, last column) closely align with the ground truth trajectories in comparison with other methods. This visual confirmation further underscores the efficacy of local exposure correction and specularity inpainting in improving the accuracy and fidelity of endoscope trajectory reconstruction within the colon compared to global gamma correction and specularity removal pipelines.

TABLE 3.1 – Quality metrics. The $\overline{APE}$ and $RMSE$ values are determined by using simultaneously all AP_i values of the four videos, while the *mean*, *std*, *median* values are computed with the $mean_j$ and $median_j$ of the AP_i values of the four video-sequences ($j \in [1,4]$). The best values are in bold.

RNN-SLAM	$APE(\downarrow)$	$RMSE(\downarrow)$	*std*	*median*	*mean*
Without exposure correction	0.972	1.914	0.391	0.817	0.887
Endo-LMSPEC	***0.927***	***1.838***	*0.442*	0.787	0.868
Recurrent Gamma Correction Daher et al. [2023]	1.037	1.991	0.412	0.762	0.939
Endo-STTN Zhang et al. [2021]	1.023	1.897	0.433	0.825	0.854

3.4.4 Qualitative Depth Map Evaluation ◂

Given that accurate depth prediction constitutes a pivotal step in the RNN-SLAM pipeline, a qualitative assessment of depth maps generated under different photometric preprocessing conditions was performed. Figure 3.6 illustrates the effect of three image enhancement strategies—Gamma correction, Endo-STTN, and Endo-LMSPEC—on the quality of the resulting depth maps for two representative colonoscopic frames. The first column shows the original RGB images, while the second column provides the corresponding depth predictions without any enhancement. The subsequent columns present the depth maps obtained after applying each of the enhancement methods as a preprocessing step.

In the absence of any enhancement (second column), the depth predictions exhibit clear signs of degradation. Underexposed and specular regions lead to substantial loss of geometric detail and increased blurring. These issues are particularly evident within the green rectangular regions, which correspond to areas affected by either illumination drop or specular saturation.

The Gamma correction method introduces some improvement in these problematic regions by normalizing overall brightness. However, the resulting depth maps still suffer from moderate blurriness and reduced contrast, indicating a limited capacity to recover fine structural information.

The application of Endo-STTN leads to further improvements. Its ability to temporally inpaint specularities across frames results in depth maps with slightly enhanced contrast and edge sharpness. Nevertheless, in both test cases, the transitions in depth remain somewhat smeared, suggesting that while temporal coherence is reinforced, spatial sharpness is not fully recovered.

The most visually compelling results are obtained when using the Endo-LMSPEC model. The depth maps in the final column of Figure 3.6 exhibit the most defined contours and richest detail in low-exposure areas. This model's pyramid-based processing and structural-aware learning allow for fine-scale correction of photometric inconsistencies without overcorrecting normally lit regions. Notably, in both examples, the anatomical boundaries within the green boxes are better preserved, with fewer artifacts and sharper gradients, leading to more reliable depth information for subsequent surface reconstruction.

These qualitative results are consistent with the quantitative improvements reported in Table 3.1,

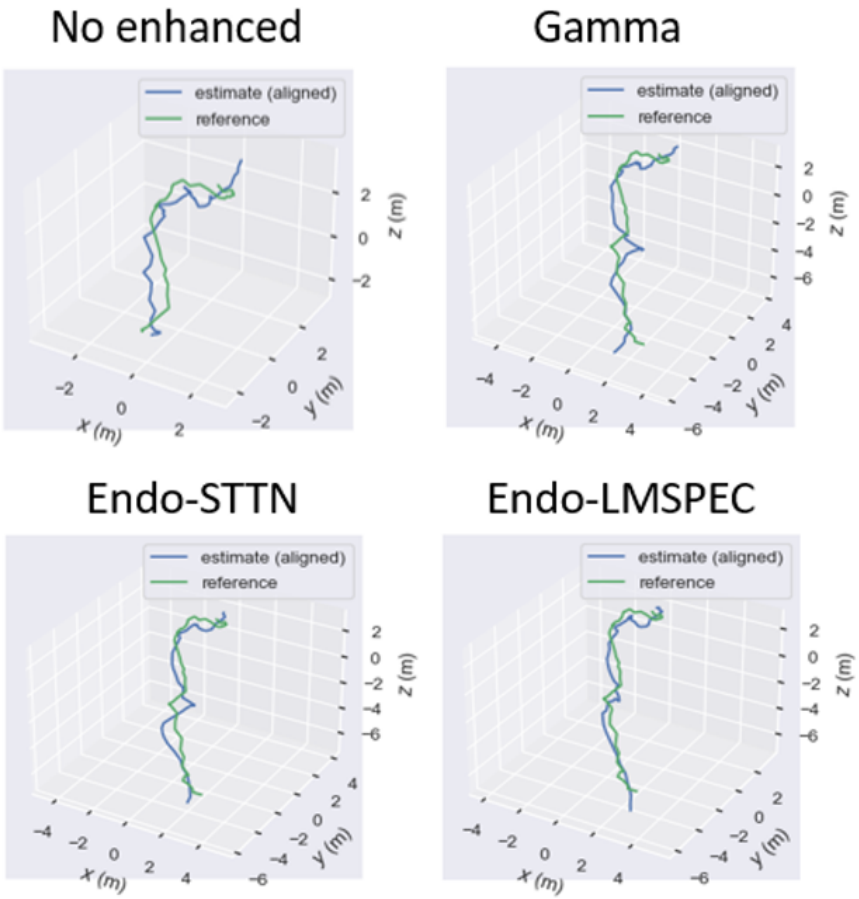


FIGURE 3.5 – Comparison of estimated camera trajectories under different image enhancement (IE) methods. The ground-truth reference trajectory is shown in green, while the estimated trajectory is shown in blue after alignment. From left to right, top to bottom: baseline without enhancement, Recurrent Gamma Correction, Endo-STTN, and Endo-LMSPEC.

confirming that Endo-LMSPEC is particularly well-suited for enhancing depth prediction robustness in monocular colonoscopic sequences characterized by challenging lighting conditions.

3.4.5 Discussion ◂

The experimental results presented in this chapter underscore the critical role of photometric pre-processing in improving the accuracy and robustness of monocular depth estimation pipelines for endoscopic 3D reconstruction. As highlighted in the trajectory evaluation (Table 3.1) and depth map visualizations (Figure 3.6), the integration of deep learning-based enhancement techniques significantly improves both pose estimation and depth prediction.

Among the enhancement strategies examined, the Endo-LMSPEC model demonstrated the most consistent improvements across all evaluation metrics. Its structural-aware, pyramid-based exposure correction proved particularly effective in restoring geometric detail in regions affected by lighting

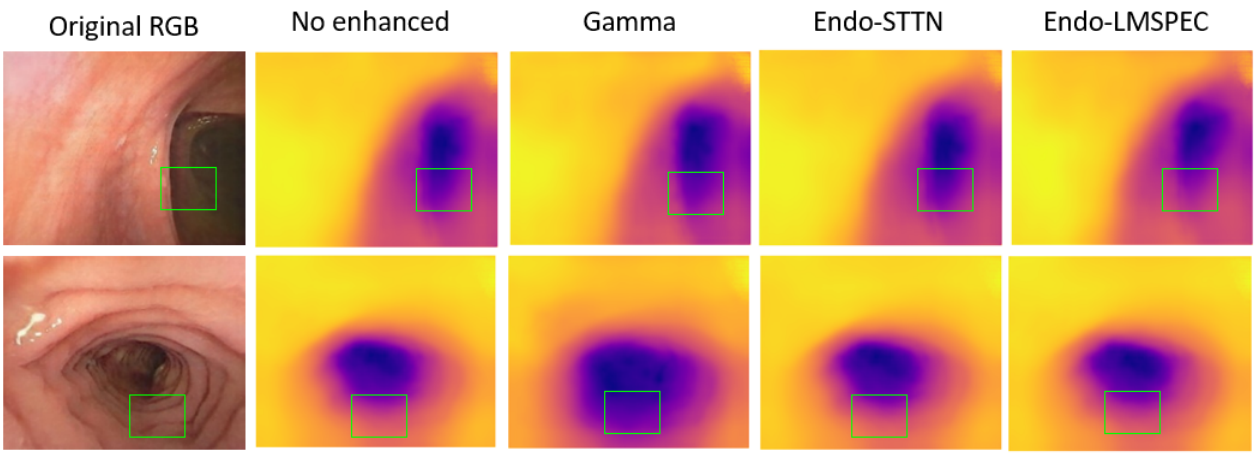


FIGURE 3.6 – Depth maps resulting from the RNN-SLAM pipeline under different image enhancement and specularity removal strategies. From left to right: Original RGB input, depth maps without enhancement, and those obtained using Gamma correction, Endo-STTN, and Endo-LMSPEC, respectively. Green boxes highlight regions with prominent photometric artifacts.

artifacts. Unlike traditional Gamma correction, which applies global adjustments, or temporal inpainting methods like Endo-STTN that target specularities, Endo-LMSPEC adapts locally to spatial illumination inconsistencies while preserving anatomical features critical for accurate depth inference. This capability was especially evident in the improved clarity and contrast of depth maps, which in turn led to lower Absolute Pose Error (APE) and Root Mean Square Error (RMSE) values.

Notably, the enhanced inputs also positively influenced the temporal coherence of the RNN-SLAM pipeline. Since recurrent depth prediction and pose refinement rely on consistent photometric patterns across frames, improvements in input quality directly impacted downstream components such as keyframe selection and local optimization. This suggests that even in the presence of robust recurrent modeling, the reliability of visual SLAM pipelines remains heavily dependent on the quality of the incoming frames.

Furthermore, these findings support the broader conclusion that self-supervised monocular reconstruction in real-world endoscopy can benefit substantially from targeted photometric pre-processing. Given the prevalence of underexposed, overexposed, and specular regions in clinical colonoscopy data, the application of exposure-aware enhancement models should be considered a necessary component of modern depth estimation frameworks.

Finally, while Endo-STTN showed notable improvements in specular region restoration, its effect on trajectory metrics was more modest, suggesting that temporal inpainting alone may not suffice when structural cues are absent or distorted. Future work may explore hybrid strategies that jointly address exposure variation and spatiotemporal inpainting to further enhance SLAM performance in endoscopic environments.

3.4.6 Conclusion ◂

This chapter presented a deep learning-enhanced 3D reconstruction pipeline for colonoscopy, emphasizing the importance of photometric pre-processing as a foundational step for improving depth and pose estimation. By integrating image enhancement techniques such as Gamma correction, Endo-STTN, and the proposed Endo-LMSPEC, the impact of illumination artifacts—specifically underexposure, overexposure, and specular reflections—was mitigated prior to the RNN-SLAM reconstruction stage.

Through both quantitative trajectory analysis and qualitative depth map inspection, it was demonstrated that high-quality image enhancement directly contributes to more stable keyframe tracking, improved depth estimation, and ultimately more accurate 3D surface reconstruction. Among the compared methods, Endo-LMSPEC emerged as the most effective strategy, offering localized correction without compromising structurally informative regions.

These findings validate the proposed approach of decoupling enhancement from reconstruction, and show that leveraging dedicated pre-processing networks as modular components can significantly boost the performance of self-supervised SLAM systems in clinical endoscopic settings. This modularity also opens the door to future integration of more advanced photometric or geometric priors, tailored for the unique challenges of real-world gastrointestinal imaging.

Publications related to chapter 3.

— **A Deep Learning-Based Image Pre-Processing Pipeline for Enhanced 3D Colon Surface Reconstruction Robust to Endoscopic Illumination Artifacts**,
Ricardo Espinosa, Carlos Axel García-Vega, Gilberto Ochoa-Ruiz, Dominique Lamarque, Christian Daul.
In: *Proceedings of the 2024 IEEE 37th International Symposium on Computer-Based Medical Systems (CBMS)*. https://ieeexplore.ieee.org/document/10234568

— **Deep Learning-Based Image Exposure Enhancement as a Pre-Processing for an Accurate 3D Colon Surface Reconstruction**,
Ricardo Espinosa, Carlos Axel García-Vega, Gilberto Ochoa-Ruiz, Dominique Lamarque, Christian Daul.
In: *Proceedings of the 2023 GRETSI Symposium on Signal and Image Processing*

Chapter 4

Illumination Invariant Self-Supervised Depth Prediction in Endoscopy

---

4.1 Importance of Depth Prediction in Endoscopic 3D Reconstruction ◂

Depth estimation is a fundamental step of the three-dimensional (3D) scene representation in medical endoscopy. Accurate depth maps are critical for a range of downstream applications, including intra-operative navigation, lesion localization, surface coverage estimation, and photometric stabilization. These capabilities are particularly important in procedures such as colonoscopy, gastroscopy, and laparoscopy, where the endoscope provides only monocular visual input and no inherent depth sensing.

In computer vision-based reconstruction frameworks, depth prediction plays a key role in enabling both Structure-from-Motion (SfM) and Simultaneous Localization and Mapping (SLAM) pipelines. In classical SfM systems, sparse 3D point clouds are generated via triangulation by solving the epipolar geometry between keyframe pairs. These reconstructions are sensitive to accurate and more or less temporally stable feature correspondences, which in turn depend on photometric and geometric consistency. In monocular SLAM, depth prediction contributes not only to static 3D map construction but also to continuous pose estimation and viewpoint tracking (Grasa et al. [2014]; Ma et al. [2019]; Phan et al. [2020]).

The potential of SLAM-based methods in endoscopic imaging has been demonstrated in methods such as RNNSLAM, where real-time 3D cartography of colon chunks is achieved using recurrent neural networks trained to predict relative poses and dense depth maps from RGB sequences (Ma et al. [2019, 2021]). These approaches rely on the accurate and temporally coherent prediction of depth, which is used to reconstruct the inner mucosal surface and maintain geometric alignment over time. In the absence of reliable depth cues, drift and misalignment often occur, particularly in regions with low texture or specular reflections.

Beyond navigation and reconstruction, dense depth maps also support the registration of endoscopic video with preoperative 3D data, the extension of fields-of-view via panoramic stitching, and the projection of semantic information into 3D space. For instance, lesion annotations can be back-projected onto reconstructed colon surfaces, aiding in localization and surgical planning. In such workflows, the quality of the initial depth estimation has a direct impact on spatial accuracy and clinical interpretability (Mahmood and Durr [2018]; Widya et al. [2019]).

Recent studies have also leveraged predicted depth for volumetric rendering and multi-view synthesis, enabling realistic simulation of camera motion and photometric behavior. These techniques facilitate the development of synthetic endoscopy scenes that are used for training, validation, and domain adaptation purposes. However, their success depend on reliable, illumination-aware depth maps that remain robust under challenging imaging conditions typical of endoscopic procedures (Recasens et al. [2021]).

In summary, monocular depth prediction has emerged as a key enabling technology in endoscopic 3D reconstruction. It contributes directly to geometric modeling, navigation, and scene understanding in a variety of clinical contexts. The following section discusses the specific challenges that hinder depth estimation in endoscopy, and sets the foundation for the self-supervised approach developed in this thesis.

4.2 General context of Depth Estimation in endoscopy ◂

4.2.1 Challenges of Depth Estimation ◂

3D reconstruction of endoscopic data with monocular depth estimation remains a challenging task due to the unique photometric and anatomical characteristics of internal organ scenes. Endoscopic videos are often characterized by non-uniform illumination, specular reflections, and featureless tissue surfaces that violate the assumptions of most traditional and learning-based depth estimation frameworks. A primary source of difficulty is photometric inconsistency. The endoscope's built-in light source generates highly directional and proximal illumination that varies with camera pose, distance to tissue, and the geometry of the organ wall. This results in frequent overexposure, shadowing, and non-Lambertian effects that reduce the reliability of photometric-based depth supervision (Ali et al. [2020]). An additional challenge is represented by texture sparsity. Inner mucosa surfaces of hollow organs are typically smooth, homogeneous in colors, and lack prominent textures or corners. In such regions, traditional and learning-based correspondence methods often struggle to establish reliable matches between frames, resulting in depth maps that are noisy, overly smooth, or geometrically inconsistent. This lack of visual cues hinders accurate surface reconstruction and camera localization. As shown in Figure 4.1, these limitations are particularly pronounced in underexposed areas and regions affected by specular reflections, where local geometric transitions are

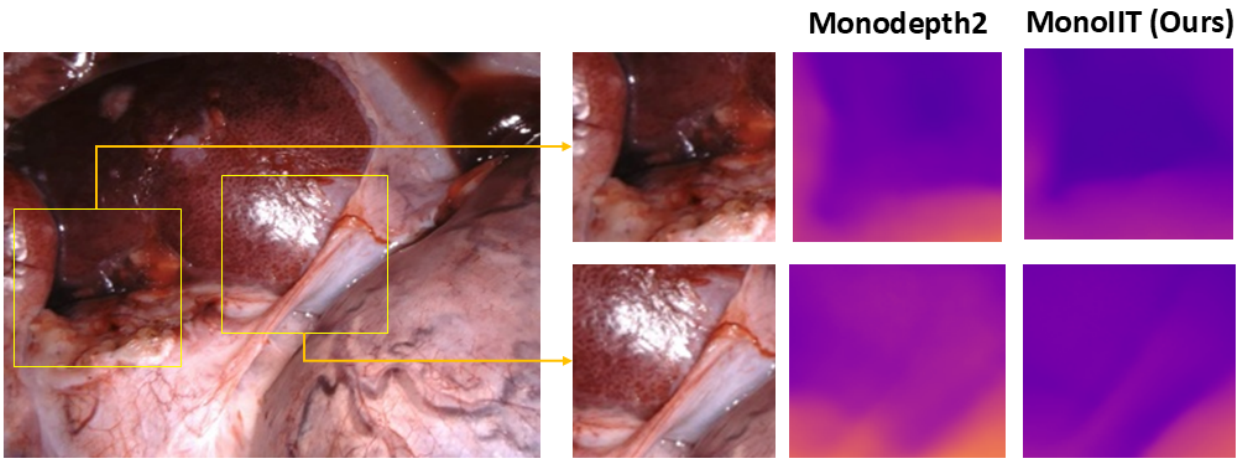


FIGURE 4.1 – Qualitative comparison of depth predictions on a representative sample from the SCARED dataset. Two regions with adverse lighting conditions, i.e., with an underexposure (top) and overexposure with specular reflection (bottom), are highlighted. Monodepth2 produces smooth but inaccurate depth estimates, while the proposed MonoIIT method recovers sharper transitions and better preserves geometric detail.

lost and depth estimates fail to align with anatomical boundaries. A third limitation stems from the absence of ground truth depth labels in real endoscopic sequences. Clinical data cannot be annotated with dense depth information due to the lack of integrated depth sensors and the impracticality of ground truth acquisition in vivo. While synthetic datasets offer partial solutions, domain shift and anatomical mismatch limit their generalization to real-world clinical settings (Mahmood and Durr [2018]; Luo et al. [2019]).

Finally, endoscopic video-sequences are often affected by motion blur, tissue deformation, and occlusions. These dynamic changes can introduce inconsistencies in geometric and photometric correspondences between frames, leading to unstable training in multi-view depth learning. In the absence of robust structural priors or illumination-invariant supervision, these limitations severely affect the accuracy and reliability of monocular depth models trained in the self-supervised setting.

These challenges motivate the development of a depth estimation pipeline that is explicitly designed to handle photometric distortions, spatial feature sparsity, and the absence of dense ground truth. The next section introduces a self-supervised transformer-based architecture that incorporates illumination-invariant constraints and long-range spatial modeling to address these limitations.

4.2.2 Principles of Learning Depth from Monocular Endoscopic Images ◂

Traditional monocular depth estimation models rely on learning a mapping $f_\theta : I \rightarrow D$ from an RGB image I to a dense depth map D. Early methods based on convolutional neural networks (CNNs) leveraged supervised losses with ground truth depth, while recent advances like Monodepth2 (Godard et al. [2019]) use self-supervision through photometric consistency. In this setting,

a depth map is used to warp neighboring frames, and the reconstruction error acts as a learning signal.

However, these approaches assume Lambertian surfaces and stable lighting—assumptions that fail in the endoscopic domain. As shown in Figure 4.1, CNN-based models such as Monodepth2 produce depth maps that are smooth but geometrically inaccurate when presented with specularities or shadows.

Transformer-based models such as MonoViT (Zhao et al. [2022]), have shown promising results in depth prediction tasks due to their capacity of modeling long-range dependencies via self-attention. These models are able to aggregate contextual information across distant regions of the image. However, their training signal for self-supervision still relies on brightness constancy assumption and it struggles with endoscopic scenes.

At the core of self-supervised monocular depth estimation lies the concept of *view synthesis*. Approaches based on this concept leverage geometric relationships between temporally adjacent frames to construct a training signal without requiring ground truth depth labels. Given a target image and one or more nearby source frames (typically adjacent in time), the model jointly predicts the depth of the target frame and the relative camera pose [1] between two views. Using these predictions, the source frame is *warped* into the target frame's viewpoint via differentiable image projection. The synthesized target frame is then compared against the original frame using a photometric loss (e.g., based on a L1-norm value or a SSIM determined for two images), effectively encouraging the network to learn geometrically consistent depth and motion estimates. This approach exploits the natural redundancy in sequential video data and forms the foundation of many recent self-supervised depth learning frameworks.

In sum, monocular depth estimation in endoscopy requires models that (i) can generalize under adverse lighting, (ii) are resilient to texture sparsity, and (iii) do not depend on ground truth depth. The next chapter introduces MonoIIT, a transformer-based self-supervised model equipped with photometric-invariant constraints to tackle these challenges.

4.3 New Self-Supervised Illumination Invariant Depth Prediction in Endoscopy ◂

4.3.1 Illumination Invariance in Endoscopy ◂

4.3.1.1 Modeling Complex Illumination Changes ◂

Ozyoruk et al. (Ozyoruk et al. [2021]) modeled illumination variations between two images using a global affine transformation defined by a multiplicative and an additive term. Although

1. A camera pose includes both the camera position and the camera orientation in a 3D coordinate system.

effective for a globally homogeneous illumination of the scene, this model is insufficient in endoscopy, where illumination changes are highly localized due to complex organ surface geometries, variations in endoscope viewpoints, and small tissue-endoscope distances. Under these conditions, a single affine transformation with constant parameters across all image regions fails to capture the complexity of the illumination variability.

A local affine model is proposed in this thesis to take into account complex illumination changes between images. Within this framework, illumination changes are modeled independently in small image regions (patches) centered at each pixel. The relationship between corresponding small patches $\mathscr{P}_s(\mathbf{x}_i)$ in source image I_s and $\mathscr{P}_t(\mathbf{x}_j)$ in target image I_t is defined by:

$$\mathscr{P}_t(\mathbf{x}_j) = \mathscr{P}_t(\mathbf{x}_i + \vec{v}) = a_{\mathbf{x}_i}\mathscr{P}_s(\mathbf{x}_i) + b_{\mathbf{x}_i}, \quad \forall \mathbf{x}_k \in \mathscr{P}_s(\mathbf{x}_i), \tag{4.1}$$

In Eq. (4.1), $a_{\mathbf{x}_i} \in \mathbb{R}_{>0}$ and $b_{\mathbf{x}_i} \in \mathbb{R}$ denote the local multiplicative and additive illumination parameters, and $\vec{v} = (x_j - x_i, y_j - y_i)$ represents the spatial displacement between the central pixels of the source ($\mathbf{x}_i = (x_i, y_i)$) and target patches ($\mathbf{x}_j = (x_j, y_j)$). As depicted in Fig. 4.2, local illuminations depend both on the local surface orientations and on the camera viewpoints (or poses).

The local affine model enables fine-grained modeling of spatially varying illumination changes across the image. Particularly when applied to small patch sizes, such as 3×3 pixels, it allows for robust handling of complex illumination effects that cannot be captured by global transformations.

4.3.1.2 Illumination-Invariant Patch Content Descriptors ◄

A key challenge in developing robust depth prediction models lies in designing features that are insensitive to the photometric changes introduced by local illumination variations. In this context, it becomes necessary to represent the content of small image regions through descriptor values that remain constant even when intensity transformations occur.

The goal, therefore, is to construct a descriptor $\mathbf{D}$ that captures the intrinsic color/grey-level information in patches while being invariant to the affine illumination changes modeled in Section 4.3.1.1. More precisely, for each patch $\mathscr{P}_s(\mathbf{x}_i)$ in the source image and its corresponding patch $\mathscr{P}_t(\mathbf{x}_j)$ in the target image, the descriptor must satisfy:

$$\mathbf{D}\left(\mathscr{P}_t(\mathbf{x}_j)\right) = \mathbf{D}\left(\mathscr{P}_t(\mathbf{x}_i + \vec{v})\right) = \mathbf{D}\left(a_{\mathbf{x}_i}\mathscr{P}_s(\mathbf{x}_i) + b_{\mathbf{x}_i}\right), \tag{4.2}$$

where $a_{\mathbf{x}_i}$ and $b_{\mathbf{x}_i}$ denote the local multiplicative and additive illumination parameters, respectively.

This property ensures that variations in lighting do not affect the matching of patches during the self-supervised training process, thereby improving the robustness of depth and pose estimation in highly variable endoscopic environments.

Inspired by the general framework proposed by Trinh et al. (Trinh and Daul [2019]), the des-

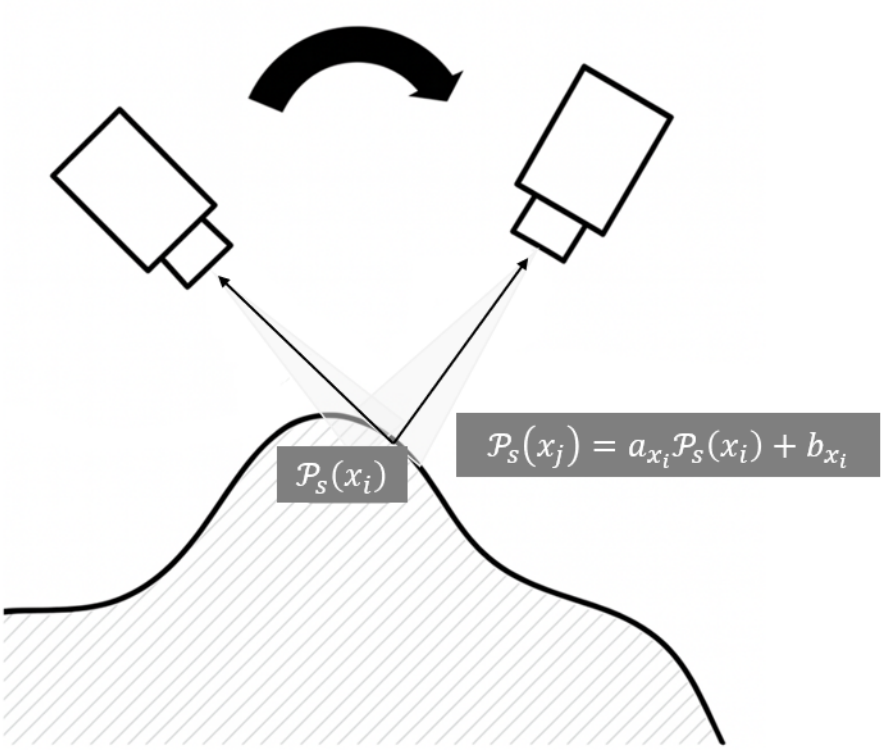


FIGURE 4.2 – Illustration of the local illumination model in endoscopic imaging. Two camera poses allow to observe the same region of an organ surface from different viewpoints. The local intensity transformation between source patch $\mathscr{P}_s(\mathbf{x}_i)$ and target patch $\mathscr{P}_t(\mathbf{x}_j)$ is modeled by a multiplicative term $a_{\mathbf{x}_i}$ and an additive term $b_{\mathbf{x}_i}$, leading to the affine relation $\mathscr{P}_t(\mathbf{x}_j) = a_{\mathbf{x}_i}\mathscr{P}_s(\mathbf{x}_i) + b_{\mathbf{x}_i}$. The values of $a_{\mathbf{x}_i}$ and $b_{\mathbf{x}_i}$ are specific and constant for the pixels of each homologous patch pair ($\mathscr{P}_s(\mathbf{x}_i)$, $\mathscr{P}_t(\mathbf{x}_j)$) seen in field of views with small apertures (gray regions in the figure).This model accounts for complex photometric variations induced by camera displacement, local surface curvature, and endoscopic lighting proximity.

criptor is constructed using directional derivatives computed via convolutional kernels. Each kernel K^d captures intensity variations along one of eight directions $d \in \{1,2,\ldots,8\}$, corresponding to two horizontal, two vertical, and four diagonal orientations:

$$K^d = \begin{bmatrix} k_1^d & k_2^d & k_3^d \\ k_4^d & k_0^d & k_5^d \\ k_6^d & k_7^d & k_8^d \end{bmatrix}. \tag{4.3}$$

Applying K^d to a patch undergoing an affine transformation yields:

$$C_{\mathbf{x}_i}^d = (a_{\mathbf{x}_i}\mathscr{P}(\mathbf{x}_i) + b_{\mathbf{x}_i}) \otimes K^d, \tag{4.4}$$

where $\otimes$ denotes the convolution operator, i.e., the sum of the products of coefficients k_m^d with $m \in$

$[0,8]$ and the nine pixel gray-level values included in patch $\mathscr{P}(\mathbf{x}_i)$ [2]. Exploiting linearity, Eq. (4.4) can be expanded as:

$$C^d_{\mathbf{x}_i} = a_{\mathbf{x}_i}(\mathscr{P}(\mathbf{x}_i) \otimes K^d) + b_{\mathbf{x}_i} \otimes K^d. \tag{4.5}$$

If the kernel is constructed such that:

$$\sum_{c=0}^{8} k^d_c = 0, \tag{4.6}$$

then

$$b_{\mathbf{x}_i} \otimes K^d = \sum_{c=0}^{c=8} b_{\mathbf{x}_i} k^d_c = b_{\mathbf{x}_i} \sum_{c=0}^{c=8} k^d_c = 0$$

and Eq. (4.5) leads to Eq. (4.7) since additive term $b_{\mathbf{x}_i}$ has no effect on the illumination change in patch $\mathscr{P}(\mathbf{x}_i)$.

$$C^d_{\mathbf{x}_i} = a_{\mathbf{x}_i}(\mathscr{P}(\mathbf{x}_i) \otimes K^d). \tag{4.7}$$

Now, one can consider vector $\mathbf{C}_{\mathbf{x}_i}$ consisting of eight components $C^d_{\mathbf{x}_i}$, i.e., $\mathbf{C}_{\mathbf{x}_i} = \left(C^1_{\mathbf{x}_i}, C^2_{\mathbf{x}_i}, \cdots, C^8_{\mathbf{x}_i}\right)^T$. The norm $\|\mathbf{C}_{\mathbf{x}_i}\|$ of this vector is defined as follows.

$$\|\mathbf{C}_{\mathbf{x}_i}\| = \sqrt{\sum_{d=1}^{d=8} \left(C^d_{\mathbf{x}_i}\right)^2} = \sqrt{(a_{\mathbf{x}_i})^2 \sum_{d=1}^{d=8} \left(\mathscr{P}(\mathbf{x}_i) \otimes K^d\right)^2} \tag{4.8}$$

The components $C^d_{\mathbf{x}_i}$ can be normalized by the norm of their vector given in Eq. (4.8).

$$\frac{C^d_{\mathbf{x}_i}}{\|\mathbf{C}_{\mathbf{x}_i}\|} = \frac{\cancel{a_{\mathbf{x}_i}}\, \mathscr{P}(\mathbf{x}_i) \otimes K^d}{\cancel{a_{\mathbf{x}_i}} \sqrt{\sum_{d=1}^{d=8} \left(\mathscr{P}(\mathbf{x}_i) \otimes K^d\right)^2}} \tag{4.9}$$

It is visible in Eq. (4.9) that this normalization compensates the illumination change effect due to multiplicative term $a_{\mathbf{x}_i}$ in patch $\mathscr{P}(\mathbf{x}_i)$. Thus, descriptor vector $\mathbf{D}$ of patch $\mathscr{P}(\mathbf{x}_i)$ can be defined as follows.

$$\mathbf{D}\left(\mathscr{P}(\mathbf{x}_i)\right) = \frac{\mathbf{C}_{\mathbf{x}_i}}{\|\mathbf{C}_{\mathbf{x}_i}\|} = \left(\frac{C^1_{\mathbf{x}_i}}{\|\mathbf{C}_{\mathbf{x}_i}\|}, \frac{C^2_{\mathbf{x}_i}}{\|\mathbf{C}_{\mathbf{x}_i}\|} \cdots, \frac{C^8_{\mathbf{x}_i}}{\|\mathbf{C}_{\mathbf{x}_i}\|}\right)^T \tag{4.10}$$

Descriptor $\mathbf{D}$ proposed in Eq. (4.10) is illumination invariant since all the components of this vector fulfill the definition of illumination invariance given in Eq. (4.2). In the specific case of oriented 3×3 derivative kernels K^d, the descriptor $\mathbf{D}$ represents the content of a patch by a star-shaped intensity variation (variations along eight star branches) which remain constant under the illumination

2. In endoscopy of hollow organs, the hue-value is almost constant over the inner epithelial surfaces so that gray-levels can be used instead of colors.

changes modeled by Eq. (4.1).

4.3.2 Towards an Illumination Invariant Self-Supervised Depth Prediction Model ◄

Recent advances in monocular depth estimation (Godard et al. [2019]; Zhou et al. [2017, 2019]; Zhao et al. [2022]) rely extensively on self-supervision signals derived from multi-view geometric constraints. In these methods, depth estimation networks are trained without explicit ground truth annotations, using instead the photometric consistency between frames acquired from different camera poses.

The supervision signal is rooted in the principle of view synthesis. Specifically, given two frames I_t (target) and I_s (source), acquired from distinct viewpoints, and assuming a pinhole camera model with known intrinsics matrix $\mathbf{K} \in \mathbb{R}^{3\times3}$, the relationship between corresponding pixels $\mathbf{x}_s^i$ and $\mathbf{x}_t^i$ can be described as:

$$\mathbf{x}_t^i = \Big[\mathbf{K} \mid \mathbf{0}\Big]\, \mathbf{M}_{s\to t} \begin{bmatrix} D(\mathbf{x}_s^i)\mathbf{K}^{-1}\mathbf{x}_s^i \\ 1 \end{bmatrix}, \tag{4.11}$$

where $\mathbf{M}_{s\to t} \in \mathbb{R}^{4\times4}$ is the 3D rigid-body transformation between the camera poses of the source and target frames, $D(\mathbf{x}_s^i)$ is the predicted depth for pixel $\mathbf{x}_s^i$ in the source frame and $\mathbf{x}_s^i$ and $\mathbf{x}_t^i$ stand for column vectors with three components representing homogeneous pixel coordinates. Under ideal conditions—accurate intrinsics, correct pose estimation, and precise depth prediction—homologous pixels are correctly aligned between I_s and I_t.

Using this geometric relationship, a synthesized target frame $\hat{I}_t$ can be generated through differentiable inverse warping (spatial transformation). Ideally, for all valid pixels i, it holds that $\hat{I}_t(\mathbf{x}_t^i) = I_t(\mathbf{x}_t^i)$.

The standard approach for quantifying appearance consistency (i.e., the similarity) between I_t and $\hat{I}_t$ involves a combination of the L_1 norm of pixel intensity differences and the Structural Similarity Index Measure (SSIM). The photometric similarity loss is defined as:

$$\mathscr{L}_{\text{PMS}}(\hat{I}_t, I_t) = \frac{\alpha(1-\text{SSIM}(\hat{I}_t, I_t))}{2} + (1-\alpha)\|\hat{I}_t - I_t\|_1, \tag{4.12}$$

where α is typically set to 0.85 to favor the SSIM component. This loss formulation balances structural and pixel-wise fidelity, ensuring that minor intensity variations do not dominate the supervision signal.

Under the assumptions of scene rigidity and viewpoint-independent illumination, the $\mathscr{L}_{\text{PMS}}$ loss provides a reliable training signal for depth and pose estimation. However, these assumptions are systematically violated in endoscopic imaging. Drastic illumination changes between frames, due to

specularities, proximity effects, and non-uniform lighting, invalidate the brightness constancy hypothesis and lead to significant depth estimation errors. In particular, regions affected by overexposure or underexposure exacerbate these issues, producing erroneous or discontinuous depth maps.

To address these challenges, the proposed method introduces two key innovations which are summarized in Fig. 4.3. First, the local affine illumination model detailed in Section 4.3.1.1 is incorporated to adapt the photometric loss $\mathscr{L}_{\text{PMS}}$ to varying lighting conditions. Specifically, terms $\|\hat{I}_t - I_t\|_1$ and $\text{SSIM}(\hat{I}_t, I_t)$ in Eq. (4.12) are corrected based on the local intensity model described in Eq. (4.1). Second, an illumination-invariant comparison strategy is employed. This strategy, introduced in Section 4.3.3.3, is based on the patch descriptors defined in Eq. (4.10). These descriptors ensure that the similarity between $\hat{I}_t$ and I_t is evaluated in a manner robust to both multiplicative and additive intensity variations.

By combining these contributions, the proposed self-supervised framework achieves significantly improved depth estimation performance in realistic illumination-challenged endoscopic scenes.

4.3.3 Training of the Proposed Architecture ◂

As depicted in Figure 4.3, during the training the proposed framework exploits three main modules: a depth estimation network, a camera pose prediction network, and a lighting correction

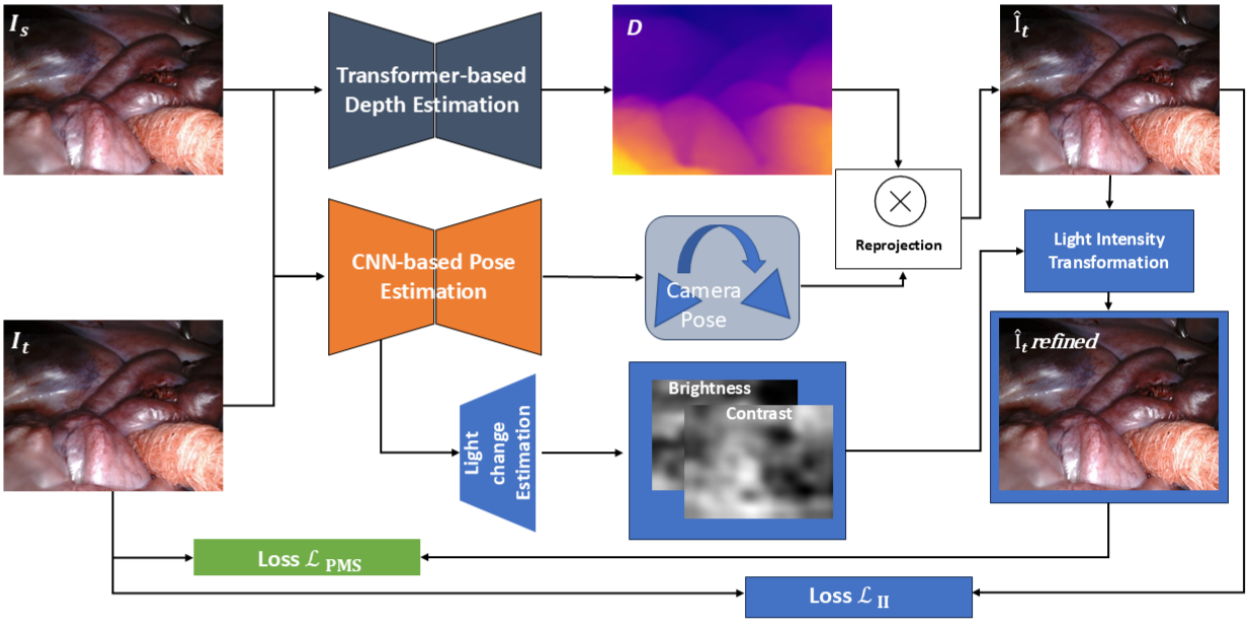


FIGURE 4.3 – Overview of the proposed training architecture. The framework integrates a depth estimation module based on CNN and transformer layers, a pose prediction module, and a lighting correction module predicting local brightness and contrast changes. Photometric and illumination-invariant losses are used to supervise the model.

calibration network.

The depth estimation module adopts a transformer-based architecture, combining convolutional and self-attention mechanisms to capture both local and global image features. The lighting correction module refines the color consistency between the synthesized target image $\hat{I}_t$ and the ground truth target image I_t by compensating for local illumination changes induced by viewpoint variations. Additionally, a novel secondary supervision signal based on an illumination-invariant loss is introduced, complementing the primary photometric loss to improve robustness against exposure artifacts.

4.3.3.1 Light Intensity Transformation ◄

The objective of the lighting change calibration module is to predict pixel-wise illumination changes between the synthesized target image $\hat{I}_t$ and the real target frame I_t.

In the affine model presented in Eq. (4.1), the multiplicative coefficient $a_{\mathbf{x}_i}$ adjusts local contrast, while the additive coefficient $b_{\mathbf{x}_i}$ corrects local brightness offsets. These parameters enable color correction across all RGB channels under the assumption that illumination variations primarily affect intensity, with negligible chromatic shifts.

As illustrated in Figure 4.4, features from the final convolutional layer of the pose encoder are decoded to produce two auxiliary outputs: a local contrast map I_A and a brightness map I_B. These maps are subsequently used to refine $\hat{I}_t$ by applying local affine corrections, aligning it closer to I_t.

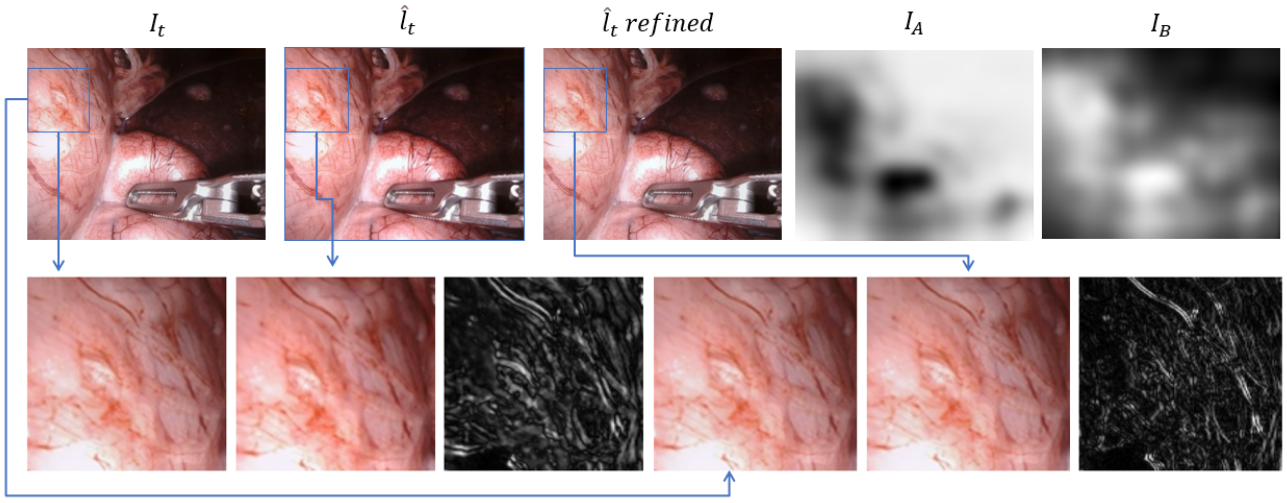


FIGURE 4.4 – Illustration of the light intensity correction process. The top row shows (left to right): target I_t, synthesized target $\hat{I}_t$, corrected synthesized image $\hat{I}_t^{\text{refined}}$, predicted contrast map I_A, and predicted brightness map I_B. The second row shows zoomed-in regions. In this row, the third and last columns respectively show the color difference maps without and and with the correction based on the illumination change prediction. The color difference image of the sixth column is with less bright values which shows a higher similarity of target I_t and the synthesized image.

4.3.3.2 Transformer-Based Depth Estimation ◄

Traditional monocular depth estimation architectures relying solely on convolutional layers Godard et al. [2019]; Ozyoruk et al. [2021]; Shao et al. [2022] often struggle to capture long-range dependencies and global spatial context. These limitations are particularly detrimental in endoscopic scenes, which are characterized by weakly textured regions, homogeneous surfaces, and low-intensity local gradients—conditions that hinder the extraction of robust geometric cues necessary for accurate depth estimation.

To address this, the proposed depth estimation module adopts a hybrid encoder-decoder architecture that integrates convolutional layers with Vision Transformer (ViT) blocks. This hybrid design enables the model to capture both local fine-grained features and global context, improving robustness in challenging visual environments.

The overall architecture, illustrated in Figure 4.5, consists of a convolutional stem followed by four hierarchical stages of "Joint CNN & Transformer Layers". Each stage includes:

— a **Multi-Scale Patch Embedding** block that extracts visual tokens using parallel convolutional filters of varying receptive fields (3×3, 5×5 and 7×7), implemented by stacking 3×3 convolutions.

— **Transformer branches** composed of multi-head self-attention layers that capture long-range dependencies across the entire feature map.

— a **Convolutional branch** that preserves locality and enhances boundary precision through 1×1 and 3×3 depthwise convolutions.

The outputs of these branches are concatenated and fused to produce rich feature representations at multiple scales. These representations are propagated through an upsampling path that uses upconvolutional layers (UpConv) and skip connections to reconstruct high-resolution depth maps.

Unlike standard Vision Transformer (ViT) models that operate on fixed-grid image patches, the adopted architecture introduces spatial flexibility, making it more suitable for the complex and irregular geometry of anatomical surfaces. By jointly encoding local spatial coherence and global contextual information, the transformer-enhanced model effectively mitigates common failure modes in depth estimation, including texture bleeding, spatial discontinuities, and oversmoothing—phenomena frequently encountered in endoscopic imaging due to the lack of strong texture and the presence of non-planar tissue structures.

This architecture is used in the MonoIIT model and was shown to significantly outperform its purely convolutional counterpart MonoII across quantitative and qualitative metrics (see Section 4.5.1 and Section 4.5.2).

Multi-scale patch embeddings and multi-branch token processing allow the network to simultaneously capture short-range and long-range dependencies, essential for depth estimation in texture-sparse endoscopic images.

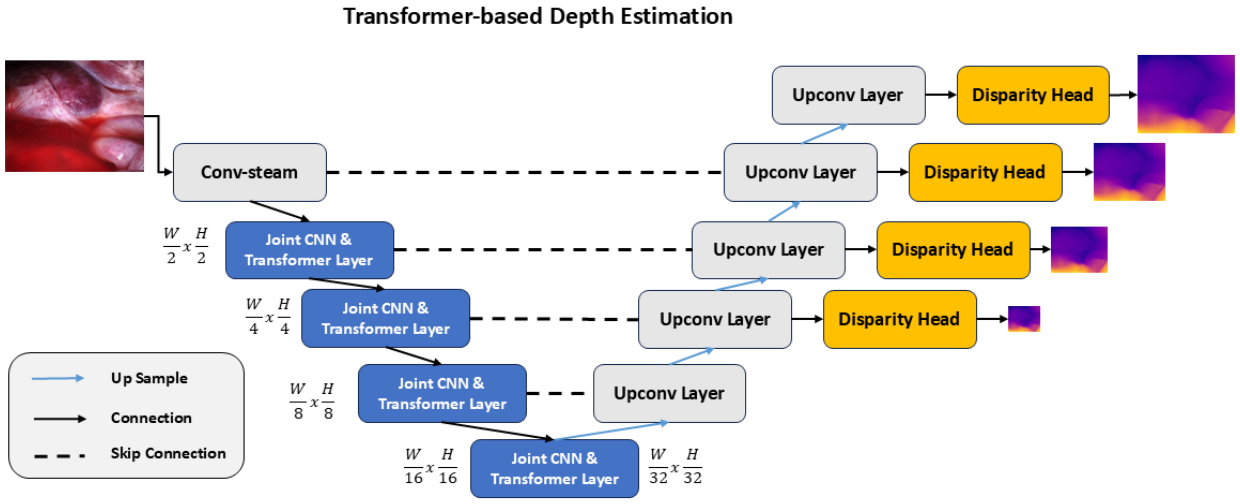


FIGURE 4.5 – Architecture of the transformer-based depth estimation framework. A convolutional stem processes the input, followed by joint CNN and transformer stages. Features are progressively downsampled and later upsampled through UpConv layers, with skip connections to preserve spatial details.

4.3.3.3 Complementary Loss Functions ◄

Overview of the Three Loss Functions.

The photometric similarity loss $\mathscr{L}_{\text{PMS}}$, defined in Eq. (4.12), constitutes the primary supervision signal for training the depth estimation network. Its effectiveness is enhanced by correcting the intensities of the synthesized target images $\hat{I}_t$ using the locally predicted multiplicative and additive parameters of the illumination model described in Eq. (4.1). However, the reliability of this correction inherently depends on the accuracy of the estimated local parameters $a_{\mathbf{x}_i}$ and $b_{\mathbf{x}_i}$ across patches $\mathscr{P}_{im}(\mathbf{x}_i)$. Inaccurate estimations may still compromise the photometric consistency signal.

An illumination-invariant loss $\mathscr{L}_{\text{II}}$ is introduced to address this limitation. This secondary supervision term is specifically designed to be independent of illumination variations due to viewpoint changes, providing a robust I_t and $\hat{i}_t$ matching signal that remains valid even under complex photometric changes.

In addition to these photometric objectives, an edge-aware depth smoothness loss $\mathscr{L}_{\text{Edge}}$ is employed. This loss encourages smooth depth variations in regions without significant image gradients (i.e., regions without edges), while preserving sharp discontinuities at anatomical boundaries, thereby ensuring the geometric plausibility of the predicted depth maps.

Illumination Invariant Loss.

The illumination-invariant loss $\mathscr{L}_{\text{II}}$ proposed in this work exploits the general formulation of patch-based illumination-invariant descriptors, as introduced in Eqs. (4.9) and (4.10). These descriptors are designed to capture local image structure independently of multiplicative and additive

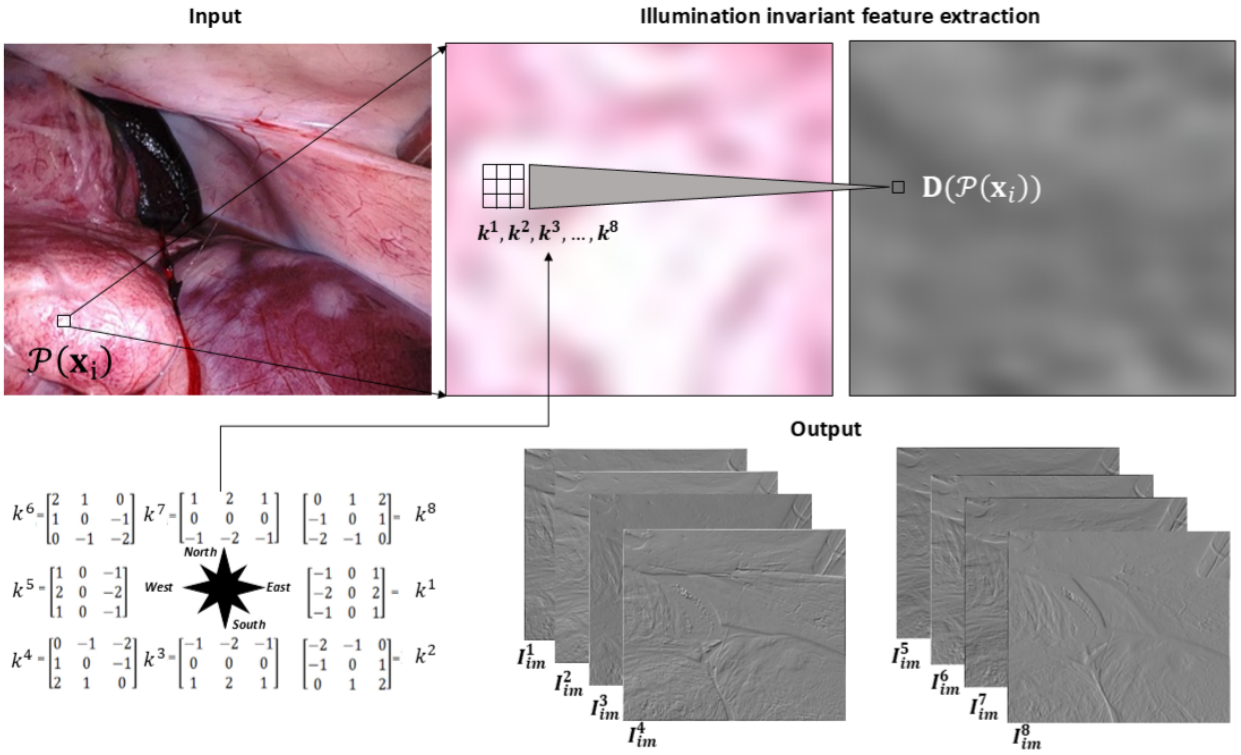


FIGURE 4.6 – Illustration of the illumination-invariant feature extraction process. Eight Robinson kernels are applied to the grayscale versions of I_s and I_t, producing directional gradient responses. After normalization (see Eq. (4.14)), the eight gradient images I^d_{im} ($d \in [1,8]$) can be used to define the illumination-invariant descriptors encoding local gray-level changes (or textures when available) independently of lighting variations.

intensity variations.

As illustrated in Figure 4.6, illumination invariance is achieved by convolving the grayscale versions of the source and target images I_s and I_t with a set of eight Robinson directional derivative kernels K^d. Each kernel corresponds to a specific orientation $d \in [1,2,\ldots,8]$. Notably, the coefficients k^d_c of each kernel are chosen such that:

$$\sum_{c=0}^{8} k^d_c = 0,$$

as indicated in Eq. (4.3), thereby eliminating the effect of the additive illumination term $b_{\mathbf{x}_i}$ in Eq. (4.5) and simplifying the convolution result to Eq. (4.7).

Following the formulation of Eq. (4.10), the illumination-invariant descriptor $\mathbf{D}^{im}(\mathscr{P}_{im}(\mathbf{x}_i))$ for each 3×3 patch $\mathscr{P}_{im}(\mathbf{x}_i)$ centered at pixel $\mathbf{x}_i$ is computed as:

$$\mathbf{D}^{im}\left(\mathscr{P}_{im}(\mathbf{x}_i)\right) = \left(\frac{C^{im,1}_{\mathbf{x}_i}}{\|\mathbf{C}^{im}_{\mathbf{x}_i}\|}, \frac{C^{im,2}_{\mathbf{x}_i}}{\|\mathbf{C}^{im}_{\mathbf{x}_i}\|}, \ldots, \frac{C^{im,8}_{\mathbf{x}_i}}{\|\mathbf{C}^{im}_{\mathbf{x}_i}\|}\right)^T \quad \text{with} \quad \|\mathbf{C}^{im}_{\mathbf{x}_i}\| = \sqrt{\sum_{d=1}^{d=8}\left(C^{im,d}_{\mathbf{x}_i}\right)^2} \tag{4.13}$$

and where $im = s$ or t indicates whether the patch is from the source or target image.

The directional responses for all patches in an image are aggregated into eight illumination-invariant (normalized) gradient maps I^d_{im}, defined by:

$$I^d_{im}(\mathbf{x}_i) = \frac{C^{im,d}_{\mathbf{x}_i}}{\|\mathbf{C}^{im}_{\mathbf{x}_i}\|} = \frac{\mathscr{P}_{im}(\mathbf{x}_i) \otimes K^d}{\sqrt{\sum_{d=1}^{8}\left(\mathscr{P}_{im}(\mathbf{x}_i) \otimes K^d\right)^2}}. \tag{4.14}$$

Once these illumination-invariant images are computed, the structural similarity index (SSIM) between the corresponding gradient maps of $\hat{I}_t$ and I_t is calculated for each direction d, and their average defines the illumination-invariant structural similarity:

$$SSIM_{\text{II}}\left(\hat{I}_t, I_t\right) = \frac{1}{8}\sum_{d=1}^{8} SSIM\left(\hat{I}^d_t, I^d_t\right). \tag{4.15}$$

Finally, the illumination-invariant loss is formulated as:

$$\mathscr{L}_{\text{II}}\left(\hat{I}_t, I_t\right) = 1 - SSIM_{\text{II}}\left(\hat{I}_t, I_t\right). \tag{4.16}$$

The illumination-invariant loss $\mathscr{L}_{\text{II}}$ acts as a complementary supervision signal in the self-supervised training framework. It is particularly beneficial in endoscopic images, which often contain large regions with weak or homogeneous texture. In these regions, standard photometric losses are highly sensitive to illumination artifacts, whereas the gradient-based descriptors enable robust matching by focusing on structural patterns independent of brightness variations.

Edge-Aware Depth Smoothness.

In the context of endoscopy, regions within an image I_{im} that exhibit minimal intensity or color variations typically correspond to geometrically smooth surfaces. Consequently, the associated depth map D should reflect these regions through small and continuous depth variations.

Formally, minor changes in pixel intensity or color around a pixel location $I_{im}(\mathbf{x}_i)$ are expected to be associated with similarly small variations in the depth map $D(\mathbf{x}_i)$. To enforce this behavior, an edge-aware smoothness loss $\mathscr{L}_{\text{Edge}}$ is adopted, as initially proposed in (Godard et al. [2019]).

In Eq. (4.17), the spatial coordinates (x_i, y_i) define the position of a pixel $\mathbf{x}_i$ within an image of dimensions $M \times N$, and the notation $\|\cdot\|$ denotes the norm of the RGB gradient vector. The partial derivatives $\frac{\partial}{\partial x_i}$ and $\frac{\partial}{\partial y_i}$ correspond to image gradients along the horizontal and vertical directions, respectively.

The edge-aware depth smoothness loss is defined as:

$$\mathscr{L}_{\text{Edge}}(D, I_s) = \frac{1}{M_{\times}N} \sum_{i=1}^{i=M\times N} \left(\left| \frac{\partial D(\mathbf{x}_i)}{\partial x_i} \right| e^{-\left\| \frac{\partial I_S(\mathbf{x}_i)}{\partial x_i} \right\|} + \left| \frac{\partial D(\mathbf{x}_i)}{\partial y_i} \right| e^{-\left\| \frac{\partial I_S(\mathbf{x}_i)}{\partial y_i} \right\|} \right), \tag{4.17}$$

where the weights $e^{-\|\cdot\|}$ modulate the importance of depth gradients based on the corresponding image gradients.

Intuitively, in regions of I_s with low intensity gradients (i.e., areas without organ borders), the exponential weighting approaches unity, thereby encouraging strong smoothing of the depth map. Conversely, near image edges (i.e., high-intensity gradients), the weighting decreases, allowing the depth map to preserve sharp discontinuities corresponding to anatomical boundaries.

Minimizing $\mathscr{L}_{\text{Edge}}$ therefore promotes spatial coherence in the depth predictions while maintaining important structural features critical for accurate endoscopic scene understanding.

4.3.3.4 Occlusion Mask Determination ◄

Due to viewpoint changes between the source and target frames, some regions that are visible in the source image I_s may become occluded in the synthesized target image $\hat{I}_t$ or the ground truth target image I_t. To account for such occlusions during loss computation, a binary occlusion mask μ is determined to identify visible and occluded pixels. The mask $\mu(\mathbf{x}_i)$ is defined such that $\mu(\mathbf{x}_i) = 1$ for visible pixels and $\mu(\mathbf{x}_i) = 0$ for occluded or unreliable regions.

The computation of the occlusion mask relies on comparing the color consistency between corresponding pixels across image pairs. Specifically, the color distance between the source and target images is defined by following norm:

$$d_{st}(\mathbf{x}_i) = \|I_s(\mathbf{x}_i) - I_t(\mathbf{x}_i)\|,$$

and the color distance between the synthesized and target images is given by:

$$d_{htt}(\mathbf{x}_i) = \left\|\hat{I}_t(\mathbf{x}_i) - I_t(\mathbf{x}_i)\right\|.$$

If the synthesized color $\hat{I}_t(\mathbf{x}_i)$ matches the target color $I_t(\mathbf{x}_i)$ more closely than the source color $I_s(\mathbf{x}_i)$ does (i.e., $d_{htt} \leq d_{st}$), the pixel is considered visible. Otherwise, the pixel is assumed to be either occluded or affected by significant photometric artifacts. Formally, the occlusion mask $\mu(\mathbf{x}_i)$ is computed as:

$$\mu(\mathbf{x}_i) = H\left(\|I_s(\mathbf{x}_i) - I_t(\mathbf{x}_i)\| - \|\hat{I}_t(\mathbf{x}_i) - I_t(\mathbf{x}_i)\|\right), \tag{4.18}$$

where $H(\cdot)$ denotes the Heaviside step function, which outputs 1 if the argument is positive and 0 otherwise.

This mask is computed dynamically during each forward pass, ensuring that loss terms such as the photometric similarity loss $\mathscr{L}_{\text{PMS}}$ and the illumination-invariant loss $\mathscr{L}_{\text{II}}$ are only evaluated over reliable, non-occluded pixels. This selective supervision significantly improves the robustness of the depth and pose learning process.

4.3.3.5 Global Loss Determination ◄

The final global loss function $\mathscr{L}_{\text{G}}$ aggregates the three complementary loss components described in the previous sections: the photometric similarity loss $\mathscr{L}_{\text{PMS}}$, the illumination-invariant loss $\mathscr{L}_{\text{II}}$, and the edge-aware smoothness loss $\mathscr{L}_{\text{Edge}}$. The photometric and illumination-invariant losses are computed selectively over non-occluded homologous pixels, identified by the occlusion mask $\mu(\mathbf{x}_i) = 1$ as defined in Eq. (4.18). This masking ensures that only reliable pixel correspondences contribute to the supervision signals. The global loss function is formally expressed as:

$$\mathscr{L}_{\text{G}}(\hat{I}_t, I_t, D, I_s) \quad = \quad \mathscr{L}_{\text{PMS}}(\hat{I}_t, I_t) + \lambda_1 \mathscr{L}_{\text{II}}(\hat{I}_t, I_t) + \lambda_2 \mathscr{L}_{\text{Edge}}(D, I_s), \tag{4.19}$$

where λ_1 and λ_2 are weighting coefficients balancing the contributions of the secondary loss terms. This formulation jointly enforces three critical objectives:

(i) The illumination corrected photometric similarity loss $\mathscr{L}_{\text{PMS}}$ encourages pixel-wise intensity consistency between the synthesized and target images (see Eq. (4.12) and Section 4.3.3.1).

(ii) The illumination-invariant loss $\mathscr{L}_{\text{II}}$, weighted by λ_1 (its value is fixed with the ablation study provided in Section 4.4), captures structural consistency based on luminance, contrast, and texture similarities, thereby enhancing robustness against severe illumination changes (see Eqs. (4.14)–(4.16)).

(iii) The edge-aware smoothness loss $\mathscr{L}_{\text{Edge}}$, weighted by a fixed λ_2 value of 10^{-3} as suggested in (Godard et al. [2019]; Shao et al. [2022]), regularizes the depth predictions by promoting spatial smoothness in textureless regions while preserving depth discontinuities at anatomical boundaries.

Together, these loss terms lead to a robust training objective that encourages photometric, structural, and geometric consistency, enabling reliable self-supervised depth prediction under challenging endoscopic imaging conditions.

4.4 Model Design and Evaluation Criteria ◄

This section first introduces the datasets employed for the model training and evaluation by detailing their acquisition protocols and characteristics. It then presents the evaluation metrics used

to quantitatively assess the model performances. The implementation details of the proposed architecture, including training of the hyper-parameters and optimization strategies, are subsequently described. An ablation study is conducted to evaluate the contribution of each component of the model and to guide the final model design. Finally, the architectures of the two proposed models are presented, emphasizing the differences between the baseline and the illumination-invariant enhanced version.

4.4.1 Depth Prediction Evaluation Datasets ◂

Three publicly available datasets were used to assess the performance of the proposed models and to enable comparisons against state-of-the-art methods.

The **SCARED** dataset consists of 35 endoscopic video-sequences acquired from fresh porcine cadaver abdominal cavities. Ground truth depth maps and ego-motion data are provided for all sequences. The dataset is divided into three subsets: 15,351 training images, 1,705 validation images, and 551 test images. The 551 test frames were employed for the ablation studies described in Section 4.4.3 and for the quantitative evaluations presented in Section 4.5, enabling a fair and systematic comparison of six different models.

The **Hamlyn** dataset (Recasens et al. [2021]) contains laparoscopic video-sequences acquired from porcine abdominal cavities, along with images of deforming silicon heart phantoms. Ground truth information is available in the form of 3D point clouds. The **SERV-CT** dataset (Edwards et al. [2022]) comprises 16 sequences of stereo image pairs captured from ex-vivo porcine torso cadavers. For the experiments presented in this work, only the left images of each stereo pair are used, along with the provided disparity ground truth. The Hamlyn and SERV-CT datasets are employed in the generalization tests described in Section 4.5, assessing the models' ability to transfer across different anatomical scenes and acquisition conditions.

The quantitative evaluation of the depth prediction accuracy was conducted using five standard error metrics, summarized in Table 4.1.

4.4.2 Performance Evaluation Protocol ◂

Comparison with Reference Models.

The performance of the proposed architecture is compared against several self-supervised monocular depth estimation baselines, including Monodepth2 (Godard et al. [2019]), MonoViT (Zhao et al. [2022]), EndoSfMLearner (Ozyoruk et al. [2021]), and AF-SfMLearner (Shao et al. [2022]). Among these, AF-SfMLearner represents the current state-of-the-art (SOTA) on the SCARED, Hamlyn, and SERV-CT datasets.

For a fair comparison, the original code provided by the respective authors is employed for Mo-

Metric	Definition
$\varepsilon_{\text{AbsRel}} =$	$\frac{1}{\lvert D_{data}\rvert}\sum_{i=1}^{\lvert D_{data}\rvert}\frac{\lvert d_i^*-d_i\rvert}{d_i^*}$
$\varepsilon_{\text{SqRel}} =$	$\frac{1}{\lvert D_{data}\rvert}\sum_{i=1}^{\lvert D_{data}\rvert}\frac{(d_i^*-d_i)^2}{d_i^*}$
$\varepsilon_{\text{RMSE}} =$	$\sqrt{\frac{1}{\lvert D_{data}\rvert}\sum_{i=1}^{\lvert D_{data}\rvert}(d_i^*-d_i)^2}$
$\varepsilon_{\text{RMSELog}} =$	$\sqrt{\frac{1}{\lvert D_{data}\rvert}\sum_{i=1}^{\lvert D_{data}\rvert}\left(\log(d_i^*+1)-\log(d_i+1)\right)^2}$
$\delta_{1.25}\ (\%) =$	$100\frac{1}{\lvert D_{data}\rvert}\sum_{i=1}^{\lvert D_{data}\rvert}H\left(1.25-\max\left(\frac{d_i^*}{d_i},\frac{d_i}{d_i^*}\right)\right)$

TABLE 4.1 – Evaluation metrics used for depth prediction assessment. Here, $|D_{data}|$ denotes the number of evaluated depths, and d_i and d_i^* represent the estimated and ground truth depth values at pixel i, respectively. The accuracy criterion $\delta_{1.25}$ uses the Heaviside function $H(\cdot)$, which outputs 1 when the maximum of $(d_i^*/d_i, d_i/d_i^*)$ is less than or equal to 1.25, and 0 otherwise.

nodepth2, MonoViT, and EndoSfMLearner. For AF-SfMLearner, evaluations are conducted using the best pretrained model publicly released by the authors.

Quality Metrics.

The models are evaluated using five commonly adopted metrics for monocular depth estimation, as summarized in Table 4.1. Here, d_i and d_i^* denote the predicted and ground truth depth values at pixel i, respectively. Specifically, d_i corresponds to the estimated depth $D(\mathbf{x}_s^i)$ obtained using Eq. (4.11).

The four error metrics employed are:

— the mean absolute relative difference ($\varepsilon_{\text{AbsRel}}$),

— the mean squared relative difference ($\varepsilon_{\text{SqRel}}$),

— the root mean squared error in linear scale ($\varepsilon_{\text{RMSE}}$) and

— the root mean squared error in logarithmic scale ($\varepsilon_{\text{RMSELog}}$).

In addition, an accuracy metric $\delta_{1.25}$ is computed, defined as the percentage of pixels for which the predicted depth falls within a $[0.8 \times d_i^*, 1.25 \times d_i^*]$ interval around the ground truth.

While the first four metrics capture the magnitude of depth estimation errors, $\delta_{1.25}$ reflects the proportion of accurate predictions within a tolerable range. It is noteworthy that the $\delta_{1.25}$ interval is asymmetric around d_i^*, favoring a tolerance of $+25\%$ and -20%. Despite the lack of a formal justification for this asymmetry in the literature, $\delta_{1.25}$ remains a gold standard metric for benchmarking monocular depth prediction methods.

Depth Normalization.

As for other monocular depth prediction methods, the outputs of the proposed model are determined up to an unknown scale factor. Consequently, during evaluation, a global scaling is applied to align the median depth of the predictions ($D_{pred}(\mathbf{x}_i)$) with that of the ground truth, following the method proposed in (Godard et al. [2019]). The scaled depth map D_{scaled} is obtained as:

$$D_{scaled}(\mathbf{x}_i) = D_{pred}(\mathbf{x}_i)\frac{\text{med}(D_{gt})}{\text{med}(D_{pred})} \quad \forall \mathbf{x}_i \in I_s, \tag{4.20}$$

where $\text{med}(\cdot)$ denotes the median operator applied to the depth values.

Following common practice (Godard et al. [2019]; Ozyoruk et al. [2021]; Shao et al. [2022]), the scaled depth maps are further capped at 150 mm for the SCARED and Hamlyn datasets, and at 180 mm for the SERV-CT dataset. These thresholds were chosen to encompass nearly all valid ground truth depth values while discarding outliers, ensuring a consistent and meaningful evaluation.

4.4.3 Ablation Study ◂

An ablation study is conducted to systematically investigate the impact of different architectural components and loss functions on the final performance. Table 4.2 summarizes the results of fourteen model variants.

The first four columns in Table 4.2 describe the configuration of each model variant:

— use (or not) of transformer-based depth estimation blocks (Section 4.3.3.2),

— inclusion (or not) of the photometric similarity loss $\mathscr{L}_{\text{PMS}}$,

— application (or not) of local lighting correction (Section 4.3.3.1),

— use and weight value of the illumination-invariant loss $\mathscr{L}_{\text{II}}$.

The remaining columns give the values of five evaluation metrics, allowing a comprehensive comparison of performance across model variants. The best and second-best results for each metric are highlighted in bold and underlined, respectively.

A series of ablation experiments were conducted on the SCARED dataset to investigate the individual contributions of different components of the proposed model. In particular, the experiments analyzed: (i) the effect of incorporating the illumination-invariant loss $\mathscr{L}_{\text{II}}$ with different weighting factors λ_1 in Eq. (4.19), (ii) the impact of local lighting correction applied within the photometric similarity loss $\mathscr{L}_{\text{PMS}}$ (weighted by $\lambda_2 = 10^{-3}$), as described in Section 4.3.3.1, and (iii) the benefit of using a transformer-based encoder-decoder architecture (Section 4.3.3.2).

Line	Transformer blocks	Use of loss $\mathscr{L}_{\text{PMS}}$	Lighting correction	Use of loss $\mathscr{L}_{\text{II}}$	$\varepsilon_{\text{AbsRel}}$ (↓)	$\varepsilon_{\text{SqRel}}$ (↓)	$\varepsilon_{\text{RMSE}}$ (↓)	$\varepsilon_{\text{RMSELog}}$ (↓)	$\delta_{1.25}$ (↑)
1	✗	✗	✗	✓($\lambda_1 = 1$)	0.072	0.713	6.027	0.098	0.942
2	✗	✓	✓	✗	0.062	0.486	5.093	0.084	<u>0.968</u>
3	✗	✓	✓	✓($\lambda_1 = 0.25$)	0.063	0.490	5.100	0.088	0.956
4	✗	✓	✓	✓($\lambda_1 = 0.5$)	0.058	<u>0.438</u>	4.850	0.082	0.966
5	✗	✓	✓	✓($\lambda_1 = 1$)	0.060	0.448	4.864	0.083	0.964
6	✗	✓	✓	✓($\lambda_1 = 2$)	0.063	0.492	5.072	0.088	0.957
7	✗	✓	✓	✓($\lambda_1 = 3$)	0.064	0.508	5.176	0.088	0.956
8	✗	✓	✓	✓($\lambda_1 = 4$)	0.064	0.506	5.193	0.088	0.960
9	✗	✓	✓	✓($\lambda_1 = 5$)	0.063	0.478	4.992	0.086	0.963
10	✗	✓	✓	✓($\lambda_1 = 10$)	0.061	0.469	4.997	0.084	0.966
11	✓	✓	✓	✓($\lambda_1 = 0.5$)	**0.055**	**0.412**	**4.614**	**0.077**	**0.969**
12	✓	✓	✓	✗	0.059	0.502	5.221	0.085	<u>0.968</u>
13	✓	✓	✗	✗	0.059	0.492	5.108	0.084	0.964
14	✓	✓	✗	✓($\lambda_1 = 0.5$)	<u>0.057</u>	0.439	<u>4.737</u>	<u>0.080</u>	0.966

TABLE 4.2 – Overview of the ablation study results. Four columns define the model configuration: i) the second column indicates whether the transformer-based depth estimation blocks described in Section 4.3.3.2 were used (symbol ✓) or not (✗), ii) the photometric similarity loss $\mathscr{L}_{\text{PMS}}$ is only used when a ✓-symbol appears in the third column, iii) ✓- marks in the fourth column indicate that loss $\mathscr{L}_{\text{PMS}}$ was computed with synthesized target images $\hat{I}_t$ using the predicted illumination change model parameters (see Section 4.3.3.1 and iv) the fifth column gives either the weight of the illumination invariant loss $\mathscr{L}_{\text{II}}$ in Eq. (4.19) or indicates that this loss in not used (i.e., symbol ✗ is equivalent to $\lambda_1 = 0$). The five last columns give the values computed for the quality criteria defined in Table 4.1. The best and second best values are respectively in bold and underlined.

Transformer Blocks.

Table 4.2 shows that models incorporating transformer-based depth estimation blocks consistently achieve lower absolute relative errors ($\varepsilon_{\text{AbsRel}}$) and squared relative errors ($\varepsilon_{\text{SqRel}}$) compared to configurations without transformers. Specifically, four out of the five best $\varepsilon_{\text{AbsRel}}$ scores are obtained with transformer blocks, regardless of whether the illumination-invariant loss $\mathscr{L}_{\text{II}}$ and/or photometric similarity loss $\mathscr{L}_{\text{PMS}}$ are used. Additionally, transformer-based models yield some of the lowest values for $\varepsilon_{\text{RMSE}}$ and $\varepsilon_{\text{RMSELog}}$, demonstrating an improved ability to minimize both global and local depth errors. These results confirm that the transformer blocks enhance the network's capacity to capture long-range dependencies and fine-grained spatial details, crucial for depth estimation in complex endoscopic environments.

Photometric Loss $\mathscr{L}_{\text{PMS}}$.

A significant improvement across all evaluation metrics is observed when the photometric similarity loss $\mathscr{L}_{\text{PMS}}$ is employed. Without $\mathscr{L}_{\text{PMS}}$ (line 1 in Table 4.2), the model exhibits the worst results across all metrics, with $\varepsilon_{\text{AbsRel}} = 0.072$, $\varepsilon_{\text{SqRel}} = 0.713$, $\varepsilon_{\text{RMSE}} = 6.027$, $\varepsilon_{\text{RMSELog}} = 0.098$, and $\delta_{1.25} = 0.942$. Introducing $\mathscr{L}_{\text{PMS}}$—either with exposure correction (line 3) or without exposure correction but in combination with $\mathscr{L}_{\text{II}}$ (line 2)—substantially improves the performance across all five metrics. These observations highlight the critical role of $\mathscr{L}_{\text{PMS}}$ as a primary supervision signal

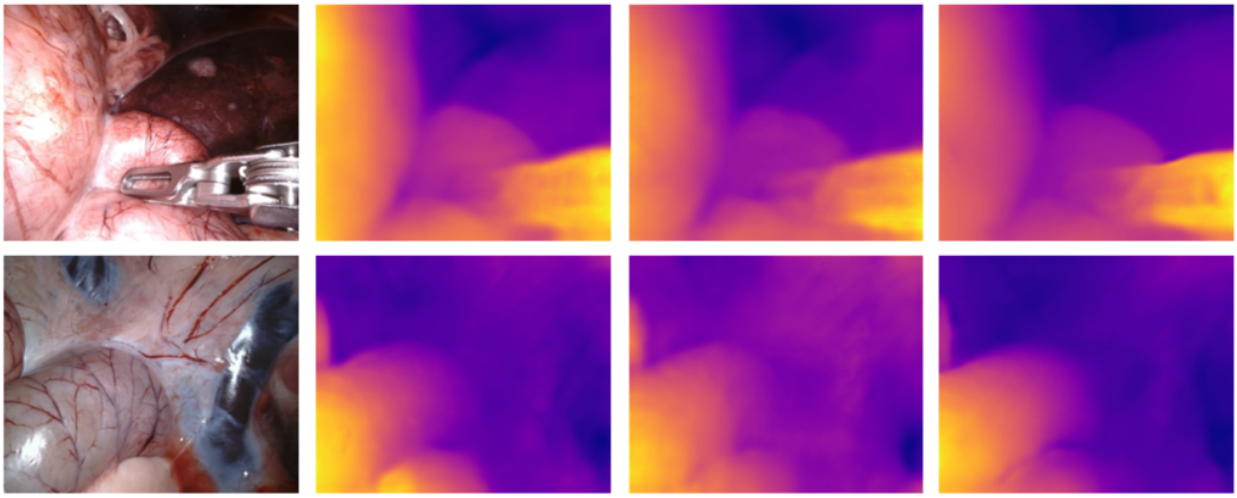


FIGURE 4.7 – Qualitative results issuing from the ablation study. The first column shows two source images I_s for which depth maps were estimated using three different loss configurations. The second column displays depth maps computed using $\mathscr{L}_{\text{PMS}}$ without exposure correction and without the illumination-invariant loss $\mathscr{L}_{\text{II}}$. The third column presents depth maps obtained with $\mathscr{L}_{\text{PMS}}$ applied after exposure correction, but still without $\mathscr{L}_{\text{II}}$. Finally, the fourth column shows depth maps estimated using both exposure-corrected $\mathscr{L}_{\text{PMS}}$ and the illumination-invariant loss $\mathscr{L}_{\text{II}}$. Progressively, from the second to the fourth column, sharper depth transitions are observed at edges in I_s, while depth variations become smoother on homogeneous surfaces, illustrating the benefits of the proposed loss design.

during training.

Illumination-Invariant Loss $\mathscr{L}_{\text{II}}$.

Models that integrate transformer blocks along with the illumination-invariant loss $\mathscr{L}_{\text{II}}$ (weighted by $\lambda_1 = 0.5$) consistently achieve the best overall performance, both with (line 11) and without (line 14) lighting correction in the photometric loss $\mathscr{L}_{\text{PMS}}$. From line 3 to line 11 in Table 4.2, the only varying factor is the value of λ_1, while the lighting correction is always active, and transformer blocks are not yet included. It is observed that setting λ_1 to 0.5 (line 4) or 1 (line 5) yields the most favorable balance across all evaluation metrics. Although a moderate degradation in performance is noted for $\lambda_1 < 0.5$ or $\lambda_1 \geq 2$, the variations remain relatively small across the range $\lambda_1 \in [0.25, 10]$. These findings indicate that while tuning λ_1 does influence performance, the model exhibits relative robustness within a specific value range. This stability simplifies the hyperparameter selection process and suggests that the method does not require precise fine-tuning to achieve reliable results.

To validate the design choices identified through the ablation study, two configurations were selected for further evaluation: one integrating transformer blocks, and one without such blocks. In both configurations, the illumination-invariant loss $\mathscr{L}_{\text{II}}$ is weighted by $\lambda_1 = 0.5$, and the photometric similarity loss $\mathscr{L}_{\text{PMS}}$ is computed using light-corrected images, as described in Section 4.3.3.1. The

benefits of combining these two supervision signals are qualitatively illustrated in Figure 4.7.

Configuration with Transformer Blocks (MonoIIT).

This configuration, corresponding to line 11 in Table 4.2, achieves the overall best performance across the five evaluation criteria. Specifically, the model obtains the lowest values for all four error metrics ($\varepsilon_{\text{AbsRel}}$, $\varepsilon_{\text{SqRel}}$, $\varepsilon_{\text{RMSE}}$, and $\varepsilon_{\text{RMSELog}}$), while the accuracy metric $\delta_{1.25} = 0.969$ ranks second, closely approaching the best value of 0.974. This configuration is referred to as **MonoIIT**, where IIT stands for "Illumination-Invariant with Transformer blocks."

Configuration without Transformer Blocks (MonoII).

This configuration, corresponding to line 4 in Table 4.2, achieves the best overall performance among the models without transformer blocks. The model ranks second according to the $\varepsilon_{\text{SqRel}}$ criterion, and achieves $\varepsilon_{\text{AbsRel}}$ (0.058) and $\varepsilon_{\text{RMSELog}}$ (0.082) values that are very close to the second-best values (respectively 0.057 and 0.080 in line 14). Additionally, its $\delta_{1.25} = 0.966$ accuracy is close to the best and second-best scores. This configuration is referred to as **MonoII**, where II stands for "Illumination-Invariant."

4.4.4 MonoIIT and MonoII Model Architectures ◂

The architectures of the MonoIIT and MonoII models are derived from the Monodepth2 baseline (Godard et al. [2019]) which is modified to enhance robustness against illumination changes through the incorporation of the $\mathscr{L}_{\text{II}}$ and $\mathscr{L}_{\text{PMS}}$ loss functions.

MonoII Model. The MonoII model adopts a conventional CNN-based design for monocular depth estimation. The depth encoder consists of a ResNet18 backbone pretrained on ImageNet (Deng et al. [2009]), while the pose estimation module is implemented as an encoder-decoder network with skip connections, also based on ResNet18 (He et al. [2016]). This architecture leverages hierarchical feature extraction to perform disparity-based depth estimation.

MonoIIT Model. The MonoIIT model architecture is inspired by the MonoViT model (Zhao et al. [2022]), integrating Vision Transformer (ViT) blocks within the depth prediction network. The architecture features a hybrid encoder-decoder structure, where ViT blocks capture long-range contextual dependencies, and convolutional layers refine local spatial details. This design, depicted in Figure 4.5, enhances the model's ability to generalize to complex, highly variable scenes by balancing local and global information. The pose estimation module employed in MonoIIT is identical to that of MonoViT.

Line	Method	$\varepsilon_{\text{AbsRel}}$ (↓)	$\varepsilon_{\text{SqRel}}$ (↓)	$\varepsilon_{\text{RMSE}}$ (↓)	$\varepsilon_{\text{RMSELog}}$ (↓)	$\delta_{1.25}$ (↑)
1	Monodepth2 (Godard et al. [2019])	0.078	0.742	6.288	0.106	0.936
2	MonoViT (Zhao et al. [2022])	0.072	0.600	5.651	0.096	0.953
3	EndoSfMLearner (Ozyoruk et al. [2021])	0.062	0.606	5.726	0.093	0.957
4	AF-SfMLearner (Shao et al. [2022])	0.059	<u>0.435</u>	4.925	<u>0.082</u>	**0.974**
5	MonoII (proposed model, first configuration)	<u>0.058</u>	0.438	<u>4.850</u>	<u>0.082</u>	0.966
6	MonoIIT (proposed model, second configuration)	**0.055**	**0.412**	**4.614**	**0.077**	<u>0.969</u>

TABLE 4.3 – Depth estimation performance evaluated on the SCARED dataset comprising 35 endoscopic video-sequences. The models were trained, validated, and tested using 15,351, 1,705, and 551 images, respectively. Best and second-best results for each metric are shown in bold and underlined, respectively.

4.5 Experimental Results ◂

This section presents a comprehensive evaluation of the proposed models. First, two quantitative experiments are described. The first evaluates performance on the SCARED dataset using the metrics defined in Table 4.1, comparing the proposed models against state-of-the-art self-supervised depth estimation methods. The second assesses the generalization ability of the models introduced in Section 4.3 on the HAMLYN and SERV-CT datasets. Finally, qualitative results are provided to visually highlight the improvements in depth estimation across endoscopic video sequences affected by illumination changes.

4.5.1 Quantitative Results ◂

4.5.1.1 Evaluation on the SCARED Dataset ◂

Table 4.3 summarizes the performance of six self-supervised depth prediction models evaluated on the SCARED dataset. A global analysis of the five evaluation metrics reveals a clear separation into two groups: a group of less performing models (lines 1 to 3) and a group of more performing models (lines 4 to 6).

The Monodepth2 model (line 1) exhibits the lowest overall performance across all five metrics. This confirms that the brightness constancy assumption made in Monodepth2 is inadequate for complex illumination conditions typically arising in endoscopy. This observation is further supported by the superior performance of models (lines 2 to 10 in Table 4.2) that incorporate illumination change modeling in either $\mathscr{L}_{\text{PMS}}$ or $\mathscr{L}_{\text{II}}$. The EndoSfMLearner (line 3) ranks as the second least performing model, indicating that a global affine correction is insufficient to fully account for the complex exposure variations encountered in endoscopy. The MonoViT model (line 2) shows slightly better performance but remains within the lower performing group, suggesting that transformer blocks alone do not sufficiently address photometric variability.

Among the better-performing models, AF-SfMLearner demonstrates significant improvements, particularly for $\varepsilon_{\text{SqRel}}$ (second best) and $\delta_{1.25}$ (best), highlighting its robustness against large depth

Line	Method	$\varepsilon_{\text{AbsRel}}$ (↓)	$\varepsilon_{\text{SqRel}}$ (↓)	$\varepsilon_{\text{RMSE}}$ (↓)	$\varepsilon_{\text{RMSELog}}$ (↓)	$\delta_{1.25}$ (↑)
1	Monodepth2 (Godard et al. [2019])	0.114	0.005	0.037	0.142	0.914
2	MonoViT (Zhao et al. [2022])	<u>0.077</u>	0.004	0.036	0.123	0.931
3	EndoSfMLearner (Ozyoruk et al. [2021])	0.169	0.013	0.063	0.246	0.733
4	AF-SfMLearner (Shao et al. [2022])	0.079	**0.003**	<u>0.032</u>	<u>0.111</u>	<u>0.958</u>
5	MonoII (proposed, first configuration)	0.080	<u>0.004</u>	0.033	0.121	0.936
6	MonoIIT (proposed, second configuration)	**0.054**	**0.003**	**0.031**	**0.098**	**0.962**

TABLE 4.4 – Performance comparison on the HAMLYN dataset. The best and second-best results for each quality metric are shown in bold and underlined, respectively.

errors. MonoII achieves comparable results, illustrating that illumination modeling can be as effective as incorporating geometric constraints such as optical flow (AF-SfMLearner). MonoIIT outperforms all models, ranking first for four metrics and second for one. The particularly low $\varepsilon_{\text{SqRel}}$ value reflects a significant reduction in large prediction errors. Comparing MonoIIT to MonoViT (both using transformer blocks) emphasizes the substantial gains achieved by explicitly correcting illumination changes through the proposed losses $\mathscr{L}_{\text{PMS}}$ and $\mathscr{L}_{\text{II}}$.

4.5.1.2 Generalization Tests ◄

The MonoII and MonoIIT models trained on the SCARED dataset were evaluated on the HAMLYN and SERV-CT datasets to assess their generalization capabilities. The corresponding results are presented in Tables 4.4 and 4.5.

In the Hamlyn dataset (Table 4.4), a similar division into two performance groups is observed. MonoViT performs slightly better than in SCARED, achieving second-best results for $\delta_{1.25}$, but the best results are again obtained by MonoIIT, ranking first across all five metrics. It can be noticed that AF-SfMLearner ranks second overall, trailing MonoIIT primarily on mean error metrics.

Similarly, in the SERV-CT dataset (Table 4.5), the MonoIIT model consistently ranks among the best-performing models, achieving three best and one second-best results across the metrics. AF-SfMLearner again ranks competitively but is slightly behind MonoIIT, while MonoVit performs relatively well within the lower performing group.

Line	Method	$\varepsilon_{\text{AbsRel}}$ (↓)	$\varepsilon_{\text{SqRel}}$ (↓)	$\varepsilon_{\text{RMSE}}$ (↓)	$\varepsilon_{\text{RMSELog}}$ (↓)	$\delta_{1.25}$ (↑)
1	Monodepth2 (Godard et al. [2019])	0.128	0.009	0.053	0.159	0.846
2	MonoViT (Zhao et al. [2022])	0.144	0.010	0.047	0.140	**0.894**
3	EndoSfMLearner (Ozyoruk et al. [2021])	0.160	0.038	0.102	0.295	0.590
4	AF-SfMLearner (Shao et al. [2022])	<u>0.108</u>	<u>0.007</u>	<u>0.046</u>	**0.136**	<u>0.882</u>
5	MonoII (proposed, first configuration)	0.115	0.009	<u>0.052</u>	0.142	0.846
6	MonoIIT (proposed, second configuration)	**0.105**	**0.006**	**0.044**	<u>0.139</u>	0.874

TABLE 4.5 – Generalization results on the SERV-CT dataset. Best and second-best results are shown in bold and underlined, respectively.

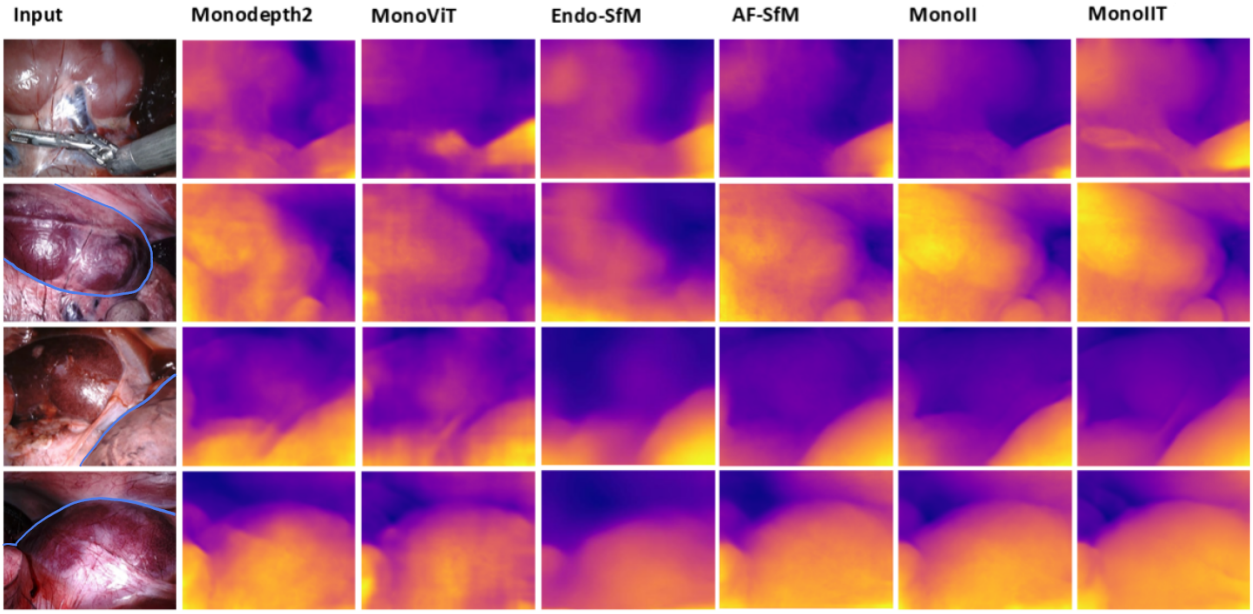


FIGURE 4.8 – Visual comparison of depth maps predicted for test images of the SCARED dataset (abdominal cavity of a pig). The first column shows, from the top to the bottom, image 264 of video sequence 6 of the dataset, image 205 of sequence 1, image 25 of sequence 5, and image 2 of sequence 3. In the depth maps (columns 2 to 7), yellow colors relate to the smallest depth values, while dark purple colors correspond to the largest depth values.

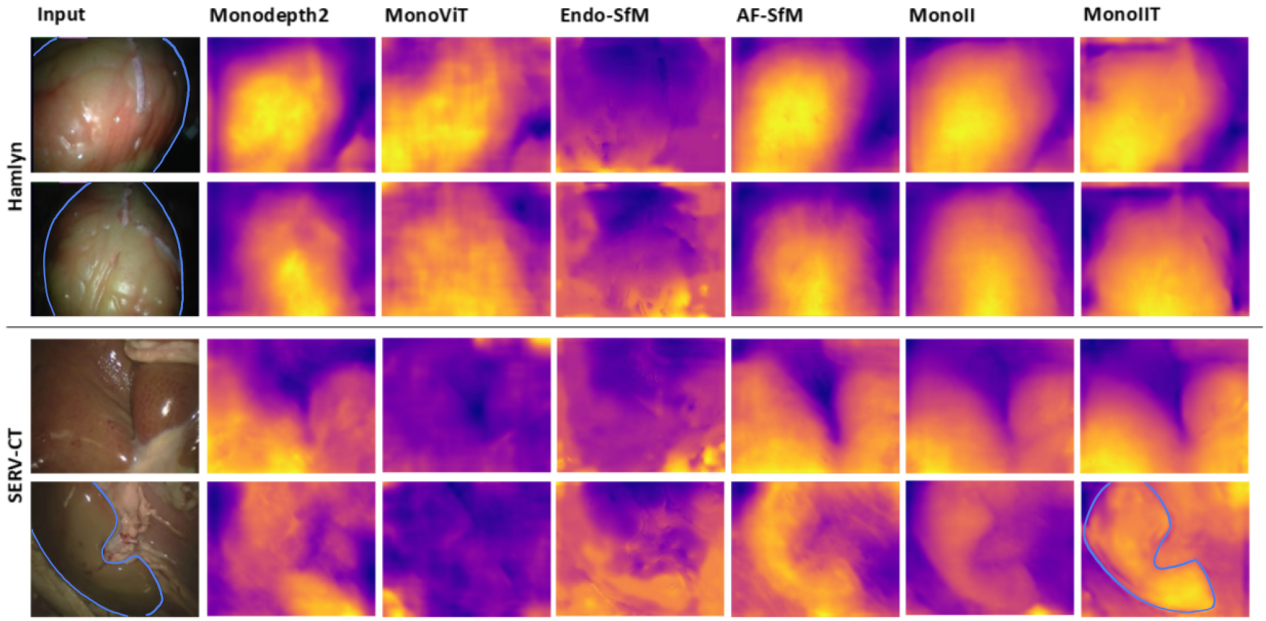


FIGURE 4.9 – Visual comparison of depth maps predicted for two images of the Hamlym and SERV-CT datasets. The Hamlyn dataset is represented by silicon heart phantom images number 0 (first line) and 1 (second line) from sequences F5 and F7, respectively. Images 001 (third line) and 0007 (fourth line) were extracted from the sequence "Experiment1" of the SERV-CT data-set (ex-vivo porcine torso cadavers).

4.5.2 Qualitative Results ◄

4.5.2.1 Visual Depth Map Comparison ◄

Figure 4.8 presents a visual analysis of the depth maps predicted for selected test images from the SCARED dataset using the six models evaluated quantitatively in Tables 4.3, 4.4, and 4.5. In the first row, the surgical tool located at the bottom right of the input image exhibits sharper edges and smoother surface representations in the MonoIIT predictions (last column). Further analysis of edge-rich regions delineated by blue curves in the second (upper left and central regions), third (lower right), and fourth rows (central and lower parts) shows that the AF-SfMLearner, MonoII, and MonoIIT models produce depth maps with the most pronounced edges and the smoothest surface transitions. Notably, the MonoIIT model yields depth maps with the strongest visual edge contrast.

Figure 4.9 extends the visual comparison to the Hamlyn and SERV-CT datasets. The first two rows correspond to Hamlyn images characterized by oval-shaped anatomical structures with smooth depth transitions. Here again, AF-SfMLearner, MonoII, and MonoIIT preserve these fine structures more accurately, with MonoIIT exhibiting the clearest and sharpest edge delineations. For the SERV-CT dataset (third and fourth rows), depth discontinuities predicted by AF-SfMLearner, MonoII, and MonoIIT align best with anatomical edges in the input images. In particular, the contrasted edge structures produced by MonoIIT are visually more accurate than those of AF-SfMLearner.

Globally, these qualitative results confirm the quantitative trends observed in Section 4.5.1. Although AF-SfMLearner and MonoIIT produce similarly coherent depth maps, MonoIIT better preserves sharp structural features and achieves higher contrast along critical anatomical boundaries.

4.5.2.2 Surface Construction ◄

Surface Construction Principle. To further assess the quality of the predicted depth maps, surface reconstructions were generated by mapping image textures onto 3D meshes created from the depth predictions. The reconstruction pipeline follows the method proposed in (Recasens et al. [2021]), which involves building a Truncated Signed Distance Function (TSDF) volume from the depth data. Each voxel in the TSDF grid stores a weighted signed distance to the nearest surface point, smoothing local noise and ensuring surface coherence.

In the case of 3D mosaicing, depth predictions from successive frames are fused into the TSDF volume to progressively build an extended surface model. Keyframes are selected to anchor depth integration, and surface meshes are extracted using the Marching Cubes algorithm (Lorensen and Cline [1987]). No global post-processing corrections are applied so that the preservation of th eraw quality of the reconstructions can visualize the intrinsic accuracy of the depth predictions.

Surface Construction from a Single Image. Figure 4.10 presents surface reconstructions from single images using depth maps generated by the AF-SfMLearner and MonoIIT models. These

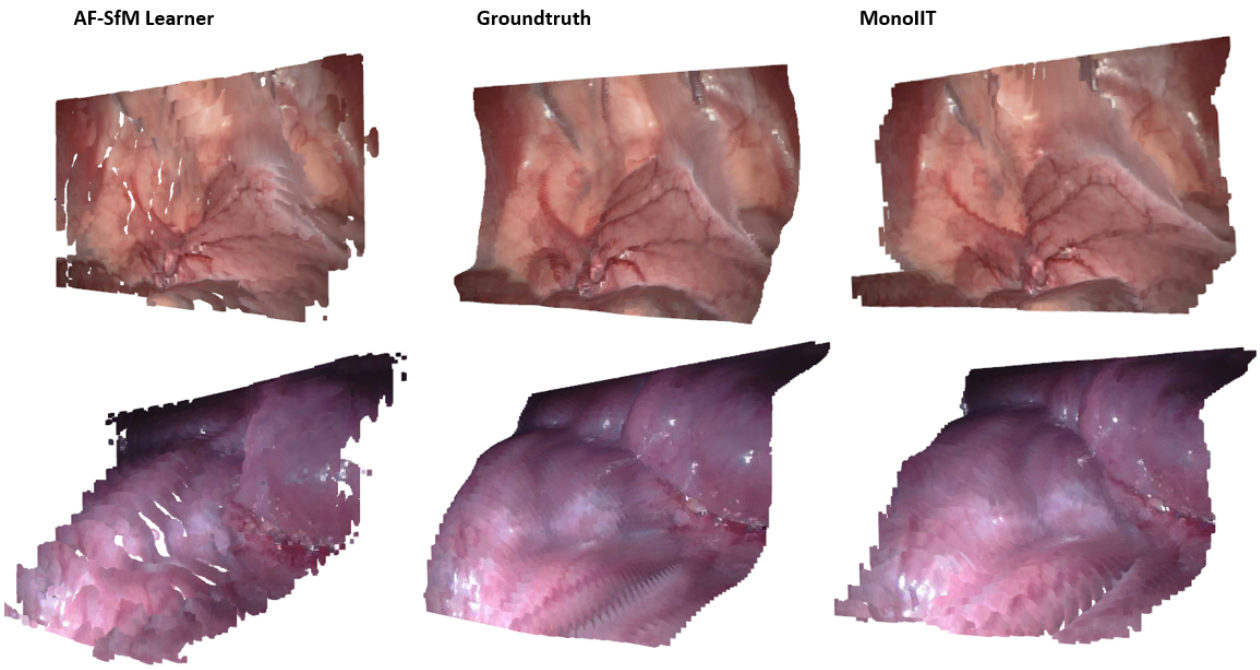


FIGURE 4.10 – Surface reconstruction results for two images from the Hamlyn dataset. The middle column shows reconstructions based on ground truth depths. The left and right columns show reconstructions using AF-SfMLearner and MonoIIT predicted depths, respectively.

surfaces can be visually compared with the reconstructions performed with the ground truth depth maps.

Surfaces generated with AF-SfMLearner (left column) exhibit notable geometric deformations, especially in regions with high curvatures as on mucosal folds. Artifacts such as staircase aliasing and over-smoothed depth transitions lead to a loss of structural details.

In contrast, MonoIIT (right column) produces surfaces that are geometrically more faithful to the ground truth. Fine structures, such as vessel patterns and localized undulations, are better preserved. Smooth gradients are maintained even in low-texture regions, and surface discontinuities are minimized. In particular, the upper row of Figure 4.10 highlights that the "tree-shaped" blood vessel structure present in the ground truth is more accurately reconstructed using MonoIIT than with AF-SfMLearner.

3D Mosaicing Test with Ten Image Sequences Figure 4.11 shows 3D mosaics constructed from sequences of ten images. The ground truth mosaic (center column) provides a reference, while the left and right columns present reconstructions exploiting AF-SfMLearner and MonoIIT depth predictions, respectively.

The AF-SfMLearner mosaic exhibits severe geometric discontinuities, particularly in regions where depth prediction errors accumulate across multiple frames. Surface parts fail to align coherently, resulting in fragmented and distorted reconstructions.

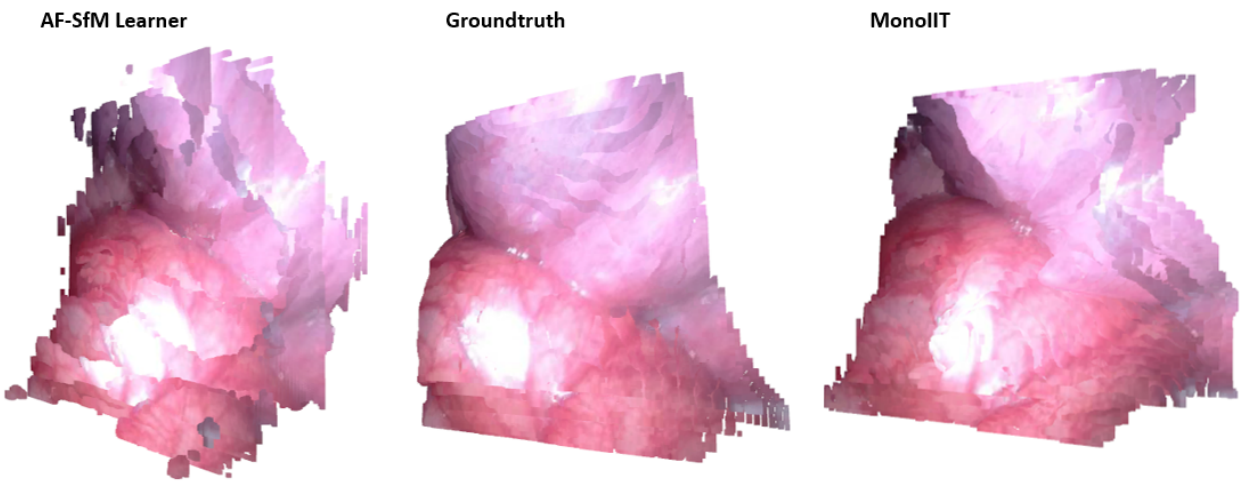


FIGURE 4.11 – 3D mosaicing results obtained using 10 consecutive frames from the Hamlyn dataset. The middle column shows the ground truth mosaic. Left: 3D mosaic obtained with the AF-SfMLearner predictions. Right: 3D map constructed with the MonoIIT predictions.

In contrast, the MonoIIT-based mosaic displays significantly improved coherence. Although some minor imperfections remain, the continuity between consecutive surface parts is largely preserved, demonstrating better cumulative consistency over sequences.

Summary.

These 3D reconstruction experiments demonstrate that although quantitative metrics suggest comparable performance between AF-SfMLearner and MonoIIT (Section 4.5.1), their practical impact on reconstructed surface quality differs significantly. While AF-SfMLearner suffers from structural inconsistencies and artifacts, MonoIIT yields geometrically faithful and visually coherent surfaces, confirming the importance of illumination modeling and transformer integration in enhancing depth estimation for endoscopic applications.

4.6 Discussion ◂

The quantitative results presented in Section 4.5.1 demonstrate that the proposed MonoIIT model achieved the highest accuracy not only on the SCARED dataset but also exhibited the best generalization performance across the Hamlyn and SERV-CT datasets. This superior performance is attributed to: (i) the integration of transformer-based depth estimation modules, (ii) the inclusion of a dedicated lighting calibration mechanism enhancing the photometric loss $\mathscr{L}_{\text{PMS}}$ (the primary supervision signal), and (iii) the introduction in the global loss of the illumination-invariant loss $\mathscr{L}_{\text{II}}$ as a complementary self-supervision signal.

AF-SfMLearner ranks as the second-best model based purely on the quantitative evaluations in Section 4.5.1. The appearance flow that this model predicts seems to offer valuable information for

depth estimation. However, as shown in Section 4.5.2.2, the two models differ substantially when used in practical textured surface reconstruction tasks. This highlights a key limitation of standard quantitative metrics: the five evaluation criteria in Table 4.1 are based on one-dimensional depth errors and do not fully capture three-dimensional spatial consistency required for extended surface reconstruction.

An important advantage of the MonoII and MonoIIT models lies in their reduced training complexity. In contrast to AF-SfMLearner, which relies on a two-stage training procedure—comprising the separate training of an external optical flow network followed by joint fine-tuning—the proposed framework adopts a single-stage, end-to-end training strategy. This unified approach eliminates dependencies on pre-trained auxiliary models, thereby improving training efficiency. When evaluated on the same hardware platform (NVIDIA DGX A100), AF-SfMLearner required approximately 9 hours and 33 minutes to complete training, while MonoIIT achieved convergence in only 5 hours and 22 minutes. This substantial reduction in training time underscores the computational advantages of the proposed method.

Despite these advances, several limitations remain, both at the depth prediction and 3D mosaicing stages.

Depth Prediction Improvements. Current challenges arise in strongly underexposed or overexposed regions, where accurate modeling of illumination changes between frames becomes difficult. In such regions, depth predictions often suffer from degraded accuracy.

One avenue for future improvement is to explicitly leverage vector fields linking homologous pixels between source I_s and target I_t. Although AF-SfMLearner attempts to use optical flow fields, their determination remain affected by complex illumination changes, limiting their effectiveness due to a lack of accuracy.

An alternative strategy would involve robust optical flow estimation methods such as those proposed in (Trinh and Daul [2019]; Weibel et al. [2012b]), which can handle large displacements, low-textured regions, and severe lighting variations. These deterministic methods could either directly align synthesized images $\hat{I}_t$ with target images I_t, or be used to supervise the training of more illumination-robust flow predictors.

3D Image Mosaicing Improvements. Current 3D mosaicing results exhibit geometric discontinuities when large sets of consecutive frames are fused without sufficient viewpoint diversity. Improvements could be achieved by strategically selecting non-consecutive frames acquired from diverse viewpoints, and by applying incremental or global bundle adjustment techniques, similar to those used in Structure-from-Motion pipelines (Phan et al. [2020]).

Such optimization techniques allow for simultaneous refinement of camera poses and surface geometry, enabling gapless and coherent surface reconstructions. Real-time applicability may be

feasible, provided that initial depth and pose estimates are sufficiently accurate.

4.7 Conclusion ◂

This work has demonstrated that the integration of an illumination-invariant loss and a per-pixel lighting calibration module within a transformer-based depth estimation framework significantly improves the robustness and accuracy of monocular depth prediction under complex illumination conditions.

The effectiveness of the proposed approach has been validated on benchmark datasets, resulting in improved depth maps and more consistent 3D surface reconstructions for endoscopic applications. The MonoIIT model consistently outperformed several state-of-the-art methods, both quantitatively and qualitatively.

Future improvements may involve the incorporation of geometric cues through illumination-invariant optical flow, aiming to further enhance the spatial coherence of depth and pose estimation. Accurate and robust depth and pose estimators are critical for the development of fully automated 3D mosaicing algorithms, not only in endoscopy but also in other domains requiring reliable scene understanding under visually challenging conditions.

While the results are promising, they also highlight the limitations of current evaluation practices. There is a growing need for standardized benchmarks that enable systematic testing under diverse real-world conditions. In this context, recent initiatives such as *EndoDepth* (Reyes-Amezcua et al. [2024]) offer an important step forward, providing tools to assess the robustness of depth prediction methods in endoscopic environments. Leveraging such benchmarks will be essential for advancing the clinical viability and

Publications related to chapter 4.

— **Transformer-Based Illumination Invariant Self-Supervised Monocular Depth Estimation in Endoscopy**
Ricardo Espinosa, Gilberto Ochoa-Ruiz, Christian Daul
Submitted to CVIU (Computer Vision and Image Understanding) on 21st March 2025 (paper under review).

— **EndoDepth: A Benchmark for Assessing Robustness in Endoscopic Depth Prediction**
Ivan Reyes-Amezcua, Ricardo Espinosa, Christian Daul, Gilberto Ochoa-Ruiz, Andres Mendez-Vazquez
In: *MICCAI Workshop on Data Engineering in Medical Imaging (DEMI 2024)*, Lecture Notes in Computer Science, Springer, 2024, pp. 84–94. First Online: 25 October 2024.

Chapter 5

3D Reconstruction of Different Internal Organs

5.1 Introduction ◂

The reconstruction of 3D anatomical structures from monocular endoscopic sequences has emerged as a powerful tool for improving navigation, scene understanding, and augmented reality in minimally invasive procedures. While significant progress has been made in monocular depth estimation, its integration into full 3D reconstruction pipelines remains limited by challenges such as poor texture, non-rigidity of organs, and illumination variability inherent to endoscopy.

This chapter extends the monocular reconstruction pipeline originally proposed in *Endo-Depth-and-Motion* (Recasens et al. [2021]) by incorporating a self-supervised depth prediction model specifically designed for endoscopic imaging, referred to as *MonoIIT*. In contrast to the original framework, which employed the Endo-Depth network for depth estimation, the present approach replaces this component with the MonoIIT model, as introduced in Chapter 4, in order to leverage its robustness to illumination variations and its improved generalization across different anatomical domains. This modification is intended to enhance the consistency of the resulting pseudo-RGBD frames and, consequently, to improve the geometric fidelity of the reconstructed 3D surfaces.

The evaluation is conducted on three representative gastrointestinal regions exhibiting increasing structural complexity and variability: the small intestine, the stomach, and the colon. For the small intestine and stomach, *ex-vivo porcine* video sequences from the EndoSLAM dataset (Ozyoruk et al. [2021]) are used, while colon reconstructions are based on *in-vivo human* recordings from (Ma et al. [2021]). This experimental setup enables a qualitative assessment of the proposed pipeline's ability to reconstruct diverse organ geometries under both controlled and clinical imaging conditions.

The contributions of this chapter include demonstrating the flexibility of MonoIIT-based depth prediction when integrated into a volumetric fusion framework, and qualitatively comparing monocular 3D reconstructions of various internal organs using a uniform pipeline. The results of this study offer insights into the applicability of depth prediction models across different anatomical areas and affirm the potential of using monocular endoscopy for extensive 3D mapping applications.

5.2 Overview of the 3D Reconstruction Pipeline ◂

The 3D reconstruction pipeline employed in this work builds upon the architecture proposed in the Endo-Depth-and-Motion framework (Recasens et al. [2021]), which enables dense mapping and camera pose estimation from monocular endoscopic sequences. The original pipeline is composed of three principal stages: (1) pixel-wise depth prediction using a supervised convolutional network; (2) Camera displacement (pose change) estimation along its trajectory between consecutive frames and keyframes; and (3) volumetric fusion of the registered depth maps using a Truncated Signed Distance Function (TSDF) representation.

The initial stage of the reconstruction pipeline is modified by replacing the Endo-Depth network with *MonoIIT*, a transformer-based model specifically designed for monocular depth prediction in endoscopic imagery. This model integrates an illumination-invariant loss function to enhance depth estimation under varying lighting conditions—an inherent challenge in endoscopic video sequences, as discussed in Chapter 4. The substitution is intended to produce more temporally consistent and anatomically accurate depth maps, thereby improving the reliability of downstream tracking and surface reconstruction stages.

The camera pose estimation stage in our pipeline follows the original implementation proposed

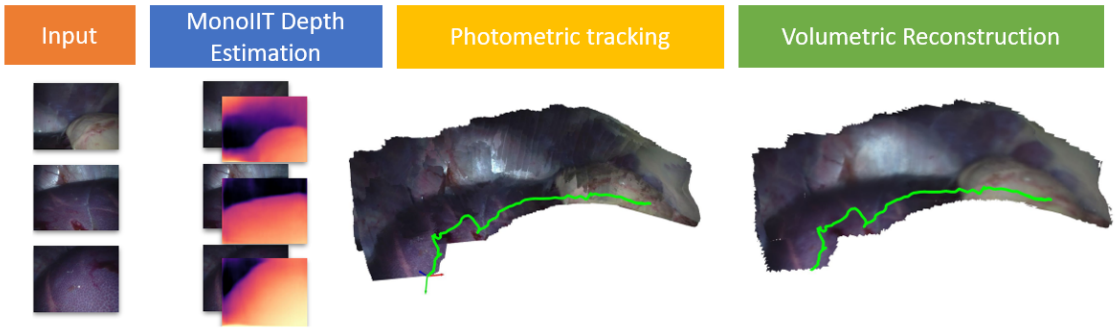


FIGURE 5.1 – Overview of the 3D reconstruction pipeline integrating MonoIIT depth prediction with photometric tracking and TSDF-based volumetric fusion.

in *Endo-Depth-and-Motion* (Recasens et al. [2021]), where pose is recovered through a photometric tracking method that estimates the rigid transformation between each frame and its nearest keyframe. Specifically, it performs a coarse-to-fine image alignment based on dense photometric residuals, optimized in a forward-compositional manner using Lie algebra updates in $\mathfrak{se}(3)$. Although it is inspired by Lucas-Kanade-style formulations, the approach operates directly on the 3D geometry provided by the predicted depth maps, not on 2D optical flow, and robustifies convergence through image pyramids and saturation of residuals. The estimated poses are used to register the pseudo-RGBD keyframes into a common reference frame. Finally, a global 3D surface is generated by integrating the aligned depth maps into a volumetric grid using Truncated Signed Distance Function (TSDF) fusion, implemented with the Open3D library (Zhou et al. [2018]).

An overview of the proposed pipeline is illustrated in Figure 5.1. The integration of MonoIIT in the depth estimation stage allows us to assess the transferability and effectiveness of this model within a real-world SLAM-like reconstruction framework. The subsequent sections detail the experimental setup and present qualitative results across multiple organ types.

5.3 Datasets and Organ Selection ◂

Three representative gastrointestinal organs were selected to evaluate the performance of our reconstruction pipeline across diverse anatomical parts: the small intestine, the stomach, and the colon. These regions were chosen due to their differing geometrical shapes, tissue appearance, and clinical relevance in diagnostic and therapeutic endoscopy.

5.3.1 Ex-vivo Porcine Data: Small Intestine and Stomach ◂

The sequences corresponding to the small intestine and stomach were obtained from the *EndoSLAM* dataset (Ozyoruk et al. [2021]), which provides high-resolution monocular endoscopic videos with precise camera pose annotations. In particular, the *High Camera Check* subset was used, which

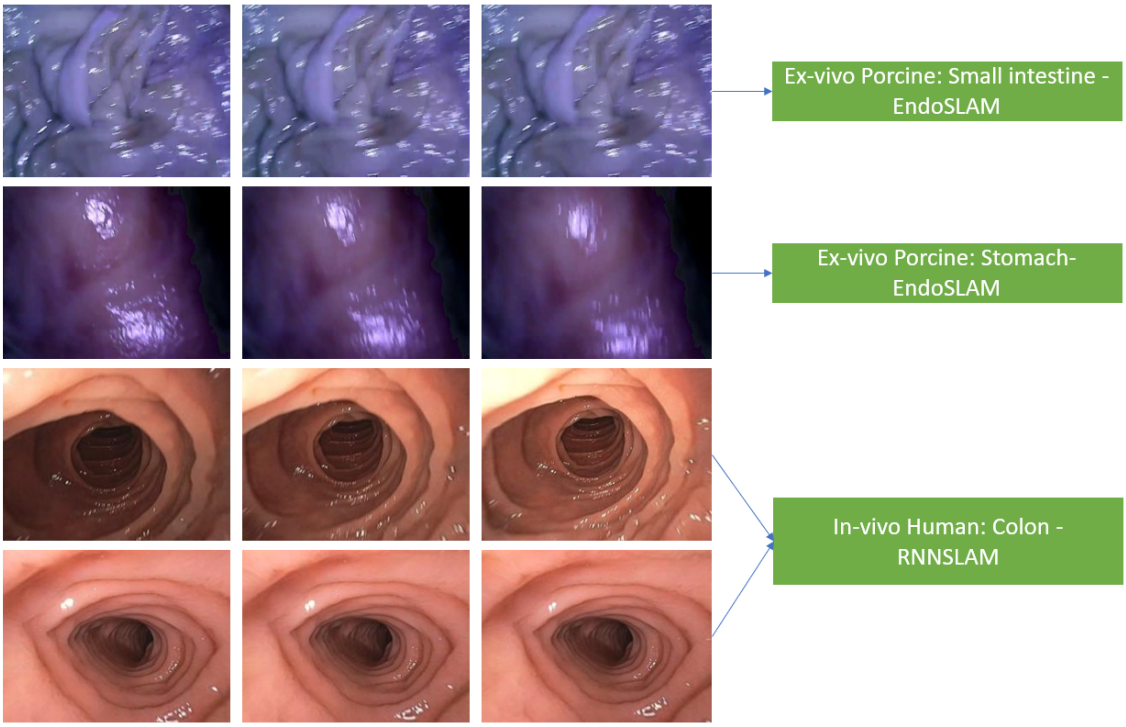


FIGURE 5.2 – Representative frames from the datasets used in our experiments. The top two rows show ex-vivo porcine sequences from the EndoSLAM dataset: small intestine (top) and stomach (middle). The bottom two rows correspond to in-vivo human colonoscopy videos from the RNNSLAM dataset. These samples highlight the variability in organ geometry, illumination, and acquisition conditions considered in the reconstruction pipeline.

contains ex-vivo porcine gastrointestinal tracts recorded under controlled experimental conditions. These sequences exhibit considerable intra-organ deformation and specularities, but benefit from a stable acquisition setup, making them suitable for evaluating the reconstruction quality under moderate complexity. The dataset provides synchronized frames and calibration files, allowing for the integration of external depth prediction models into reconstruction pipelines.

5.3.2 In-vivo Human Data: Colon ◂

To assess the performance of the proposed method in a real clinical context, in-vivo colonoscopic data from the RNNSLAM benchmark (Ma et al. [2019]) was incorporated, which focuses on reconstructing the 3D structure of the colon to identify unobserved regions during routine colonoscopy. This dataset contains monocular RGB videos acquired from live patients, and poses unique challenges including rapid camera motion, fluid presence, tissue deformation, and illumination variability. Despite the absence of depth ground truth, this dataset enables qualitative validation of depth-based 3D mapping techniques in realistic conditions.

Figure 5.2 illustrates representative frames from each organ dataset used in our evaluation. By

testing on both ex-vivo and in-vivo conditions, we aim to validate the generalizability and robustness of our pipeline across a wide spectrum of anatomical and operational scenarios.

5.4 Depth Prediction Using MonoIIT ◂

The first stage of the proposed 3D reconstruction pipeline involves predicting dense depth maps from monocular endoscopic images. For this task, *MonoIIT*, a self-supervised transformer-based model specifically designed for monocular depth estimation in medical endoscopy is used. MonoIIT introduces an illumination-invariant loss function that enhances depth predictions in the presence of severe lighting fluctuations—one of the principal challenges in gastrointestinal imaging.

Unlike conventional convolutional networks, MonoIIT leverages a hybrid encoder-decoder architecture with attention mechanisms that improve the global coherence of depth predictions, particularly in textureless or specular regions. The model is trained using a self-supervised learning paradigm, exploiting geometric constraints from monocular video sequences without requiring external depth sensors or annotated ground truth. The training objective combines photometric consistency across temporally adjacent frames with regularization terms to encourage smoothness while preserving anatomical boundaries.

In the context of the proposed experiments experiments, a pretrained MonoIIT model on endoscopic sequences from SCARED dataset was used. The model receives as input single RGB frames and outputs dense depth maps at the original image resolution. These depth maps serve as the basis for generating pseudo-RGBD keyframes, which are subsequently used for camera pose estimation and volumetric fusion.

Figure 5.3 shows representative depth predictions produced by MonoIIT for each of the organs included in our study. As observed, the model is able to capture the coarse geometry of the scene and is robust to typical endoscopic artifacts such as specular highlights, partial occlusions, and tissue folds. Notably, depth continuity is well preserved even in challenging in-vivo colon data, demonstrating the model's capacity to generalize across different anatomical domains and imaging conditions.

The integration of MonoIIT into the pipeline not only enables robust depth estimation but also aligns with the constraints of real-world clinical deployment, as it eliminates the dependency on stereo setups or preoperative scans. The resulting pseudo-RGBD frames thus provide a reliable input for downstream tracking and mapping tasks.

5.5 Camera Tracking and Volumetric Fusion ◂

Following the generation of pseudo-RGBD keyframes via MonoIIT, the next stages of the reconstruction pipeline focus on estimating camera motion and incrementally building a dense 3D

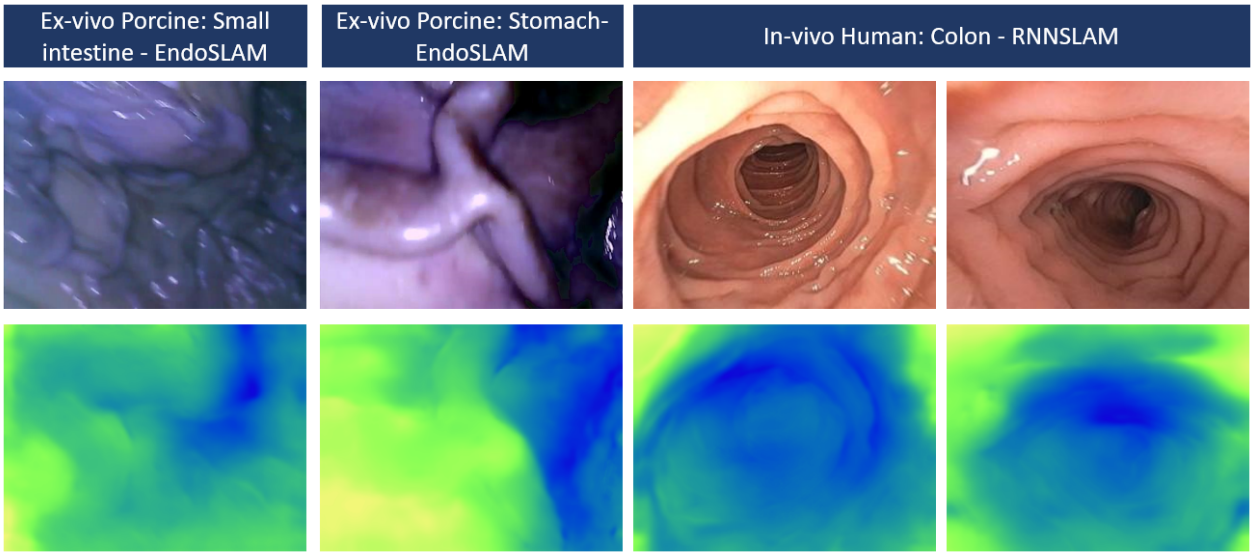


FIGURE 5.3 – Sample depth predictions generated by MonoIIT across different gastrointestinal organs. Top row: RGB input frames from the EndoSLAM dataset (small intestine and stomach, ex-vivo porcine) and the RNNSLAM dataset (colon, in-vivo human). Bottom row: corresponding depth maps predicted by MonoIIT. The model captures coarse geometry and lumen structure, exhibiting robustness to specular reflections, folds, and lighting variations.

map of the observed organ surfaces. The photometric tracking and TSDF-based volumetric fusion strategy proposed in the original Endo-Depth-and-Motion framework (Recasens et al. [2021]) was used, with no modifications to these components to ensure a fair integration of MonoIIT's depth predictions.

5.5.1 Photometric Camera Tracking ◂

Camera pose estimation is performed using a direct, keyframe-based method that minimizes photometric error between consecutive frames and their corresponding keyframes. This approach relies on the assumption of brightness constancy: a 3D point in the scene, when projected into multiple views, should have consistent intensity values across frames. Instead of relying on sparse feature correspondences, the method uses dense pixel-wise intensity residuals, enabling more accurate and stable motion estimation in texture-limited endoscopic imagery.

Given a dense depth map $D(\mathbf{x}_s)$ predicted for a source frame I_s, the transformation $\mathbf{M}_{s\to t} \in SE(3)$ from the source frame I_s to a target frame I_t is estimated by minimizing the photometric error between the target image I_t and a synthetically warped version of the source image I_s. The warping is defined by back-projecting each pixel $\mathbf{x}_s^i$ from I_s into 3D using its depth $D(\mathbf{x}_s^i)$, transforming the resulting 3D point using $\mathbf{M}_{s\to t}$, and projecting it back into the image plane of I_t using the camera intrinsics $\mathbf{K}$ as follows:

$$\mathbf{x}_t^i = \Big[\mathbf{K} \mid \mathbf{0}\Big]\, \mathbf{M}_{s\to t} \begin{bmatrix} D(\mathbf{x}_s^i)\mathbf{K}^{-1}\mathbf{x}_s^i \\ 1 \end{bmatrix}, \tag{5.1}$$

where $\mathbf{x}_t^i$ is the projection of pixel $\mathbf{x}_s^i$ from the source frame I_s into the target frame I_t, using its estimated depth $D(\mathbf{x}_s^i)$ and the relative camera transformation $\mathbf{M}_{s\to t} \in SE(3)$. The term $D(\mathbf{x}_s^i)\mathbf{K}^{-1}\mathbf{x}_s^i$ corresponds to the back-projection of the pixel into 3D camera coordinates. This 3D point is then transformed to the target frame using $\mathbf{M}_{s\to t}$, and projected back onto the image plane through the camera intrinsics using the matrix $\Big[\mathbf{K} \mid \mathbf{0}\Big]$. The synthesized target image $\hat{I}_t$ is obtained by determining the colors of its pixels at location $\mathbf{x}_t^i$ via a bilinear interpolation of the colors around pixel $\mathbf{x}_s^i$ in source image I_s.

The camera pose is optimized by minimizing the sum of robust photometric residuals:

$$\psi^* = \arg\min_{\psi} \sum_{\mathbf{x}_t^i \in \Omega_t} \rho\left(I_t(\mathbf{x}_t^i) - \hat{I}_t(\mathbf{x}_t^i)\right) \quad \text{with} \quad \Omega_t = I_t \cap \hat{I}_t, \tag{5.2}$$

and where ψ encodes the incremental pose update in the Lie algebra $\mathfrak{se}(3)$, Ω_k denotes the set of valid pixels in the keyframe, and $\rho(\cdot)$ is Huber loss function used to suppress the influence of outliers due to occlusions or non-Lambertian reflections. In the implementation, a truncated ℓ_2 loss with a threshold γ is used:

$$\rho(x) = \min(\|x\|_2, \gamma). \tag{5.3}$$

A coarse-to-fine image pyramid is employed, allowing the optimization to recover both large and small motions and thus improving convergence. At each pyramid level, the Gauss-Newton algorithm is used to iteratively update the pose parameters by linearizing the residuals and solving for the optimal increment in $\mathfrak{se}(3)$.

This direct tracking formulation offers several advantages for endoscopic reconstruction. First, it leverages the full image information, making it suitable for low-texture or specular environments where feature-based methods may fail. Second, by combining accurate depth estimates from Mono-IIT with robust photometric alignment, it provides reliable frame-to-frame motion estimation that serves as the backbone of the volumetric fusion process. Nevertheless, the method is still sensitive to significant illumination changes, rapid motion, or strong deformations, which may lead to drift or tracking failure in challenging regions.

5.5.2 Volumetric Fusion ◂

Once frame-to-keyframe poses are estimated, the corresponding depth maps are registered into a common coordinate system and fused into a volumetric representation using a Truncated Signed Distance Function (TSDF). Each voxel in the grid accumulates signed distance measurements from

the depth maps, which are integrated through a weighted average. This allows for robust aggregation of noisy predictions and smooth surface reconstruction.

The final 3D mesh is extracted using the Marching Cubes algorithm and rendered using the Open3D library (Zhou et al. [2018]). The reconstructed surfaces provide a globally consistent representation of the anatomical cavity, suitable for clinical visualization, coverage analysis, or virtual navigation.

5.6 Qualitative Results ◂

A series of experiments using representative sequences from each anatomical region were conducted to qualitatively assess the performance and visual quality of the proposed 3D reconstruction pipeline across different organ types. For each selected video, the scene was reconstructed under two configurations: (i) using a short window of 10 consecutive frames, and (ii) using the entire available sequence. This experimental design allowed to visually assess the contributions of temporal integration to surface completeness and spatial coherence.

5.6.1 Small Intestine (Ex-vivo Porcine) ◂

For the small intestine experiment, a representative ex-vivo porcine sequence from the *EndoSLAM High Camera Check* subset was selected. This sequence features forward navigation through a highly folded, low-texture lumen with prominent specularities.

Figure 5.4 illustrates the sequence structure and corresponding reconstruction outputs. The top row shows a selection of four consecutive RGB frames that highlight the appearance variability and geometric complexity of the scene. The bottom row presents the 3D surfaces reconstructed using 10 frames (left) and the full sequence (right).

The 10-frame reconstruction captures only a partial and fragmented view of the folded mucosa, with several discontinuities and holes. In contrast, the full-sequence reconstruction results in a denser and smoother surface, showing improved fold continuity, reduced fragmentation, and overall geometric coherence. These results confirm the benefit of temporal integration for accumulating depth evidence and stabilizing pose tracking in challenging anatomical regions.

5.6.2 Stomach (Ex-vivo Porcine) ◂

Two sequences from the ex-vivo porcine stomach subset of the EndoSLAM dataset were selected to assess the robustness of the reconstruction pipeline under variable inner organ shapes. These examples represent different imaging challenges due to surface reflectivity, organ curvature, and camera dynamics.

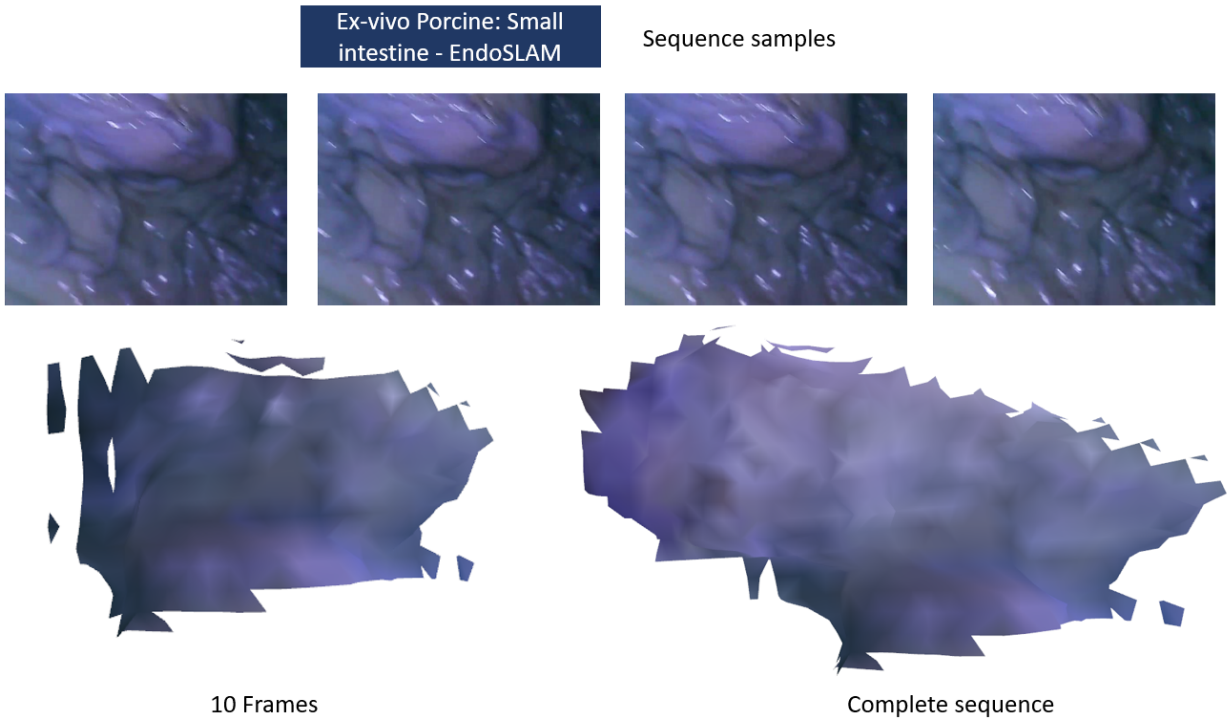


FIGURE 5.4 – Qualitative reconstruction results for an ex-vivo porcine small intestine sequence from the EndoSLAM dataset. Top: sample RGB frames from the input sequence. Bottom: reconstructed 3D surfaces using 10 frames (left) and the complete sequence (right). As the number of integrated frames increases, the reconstruction exhibits greater spatial coherence, geometric completeness, and surface continuity.

The first sequence (Figure 5.5) features a close-up view of the gastric mucosa under severe lighting variations and specularities. The top row shows RGB frames from the sequence, while the bottom row presents the reconstructions using 10 frames (left) and the complete sequence (right). The 10-frame result offers only a partial and smoothed surface, with limited geometry. The full-sequence reconstruction yields a more complete stomach wall structure, albeit with some residual deformation artifacts.

The second sequence (Figure 5.6) involves traversal over a wide field of view with more distinct mucosal folds. The 10-frame reconstruction is sparse and fragmented due to discontinuities in camera tracking and unmodeled deformations. When using the full sequence, the fusion process integrates a larger portion of the anatomy, improving coverage and producing a more plausible geometry. Nevertheless, irregularities at the boundaries remain due to motion blur and transient lighting changes.

Overall, these results illustrate the pipeline's ability to adapt to heterogeneous stomach regions and the importance of sufficient temporal integration to overcome local ambiguities and accumulate coherent structural information.

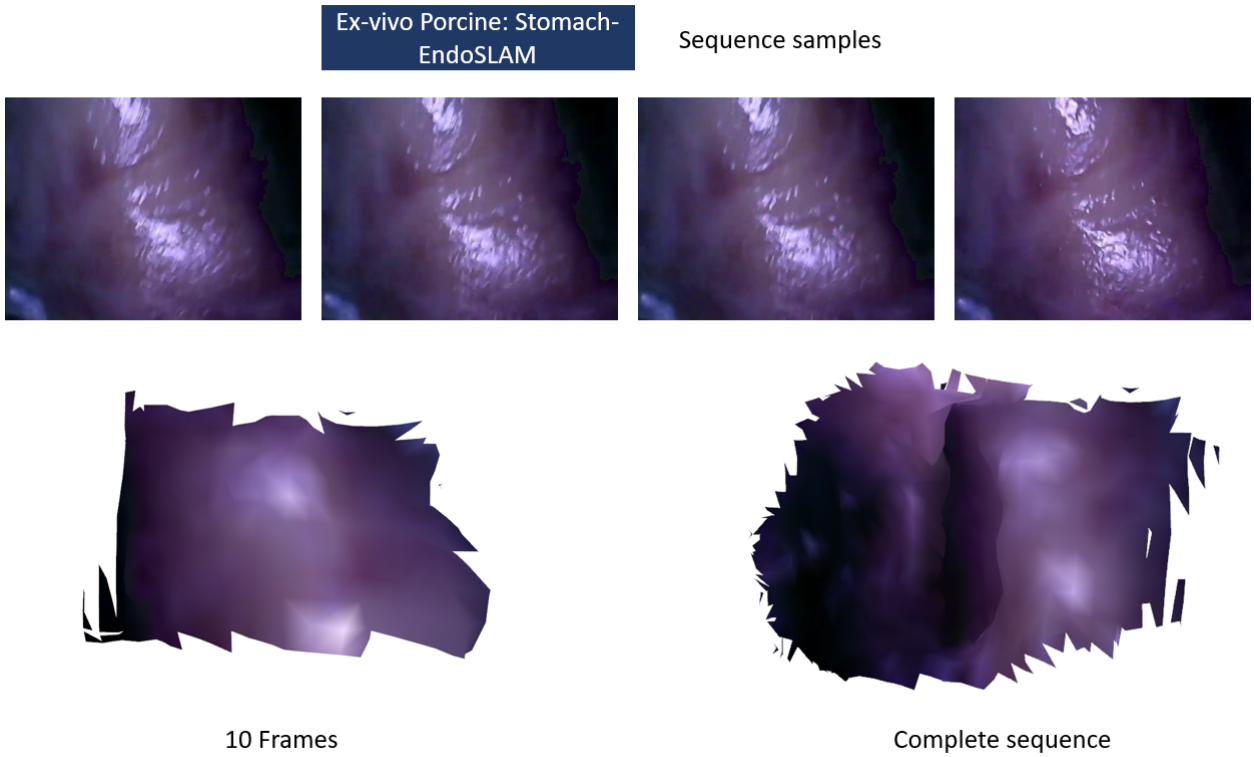


FIGURE 5.5 – Qualitative reconstruction results for an ex-vivo porcine stomach sequence from the EndoSLAM dataset. Top: sample RGB frames illustrating specularities and low-texture regions. Bottom: 3D reconstructions obtained using 10 frames (left) and the complete sequence (right). The full-sequence reconstruction exhibits improved surface coverage and geometric continuity, while the 10-frame result is limited in extent and smoothness due to insufficient integration.

5.6.3 Colon (In-vivo Human) ◂

To evaluate the reconstruction pipeline in real-world clinical settings, two in-vivo colonoscopy sequences from the RNNSLAM dataset were further tested. These sequences capture typical challenges of human endoscopic procedures, including specular reflections, peristaltic motion, fluids, and rapid changes in camera trajectory.

The first sequence (Figure 5.7) features a relatively stable camera trajectory through a well-illuminated region of the colon. The top row shows representative RGB frames from the sequence, while the bottom row displays the reconstructed 3D surfaces using 10 frames (left) and the complete sequence (right). The 10-frame result captures only a small portion of the lumen, with minor distortions and gaps. In contrast, the full-sequence reconstruction extends the mapped region substantially, recovering global colon morphology and smooth surface continuity despite specular highlights.

The second sequence (Figure 5.8) involves more complex motion and lighting variability. The 10-frame reconstruction (bottom-left) again shows partial coverage, limited by insufficient overlap and tracking noise. The complete sequence (bottom-right) yields a larger and more continuous reconstruction, capturing both the lumen curvature and wall texture. Some artifacts persist due to

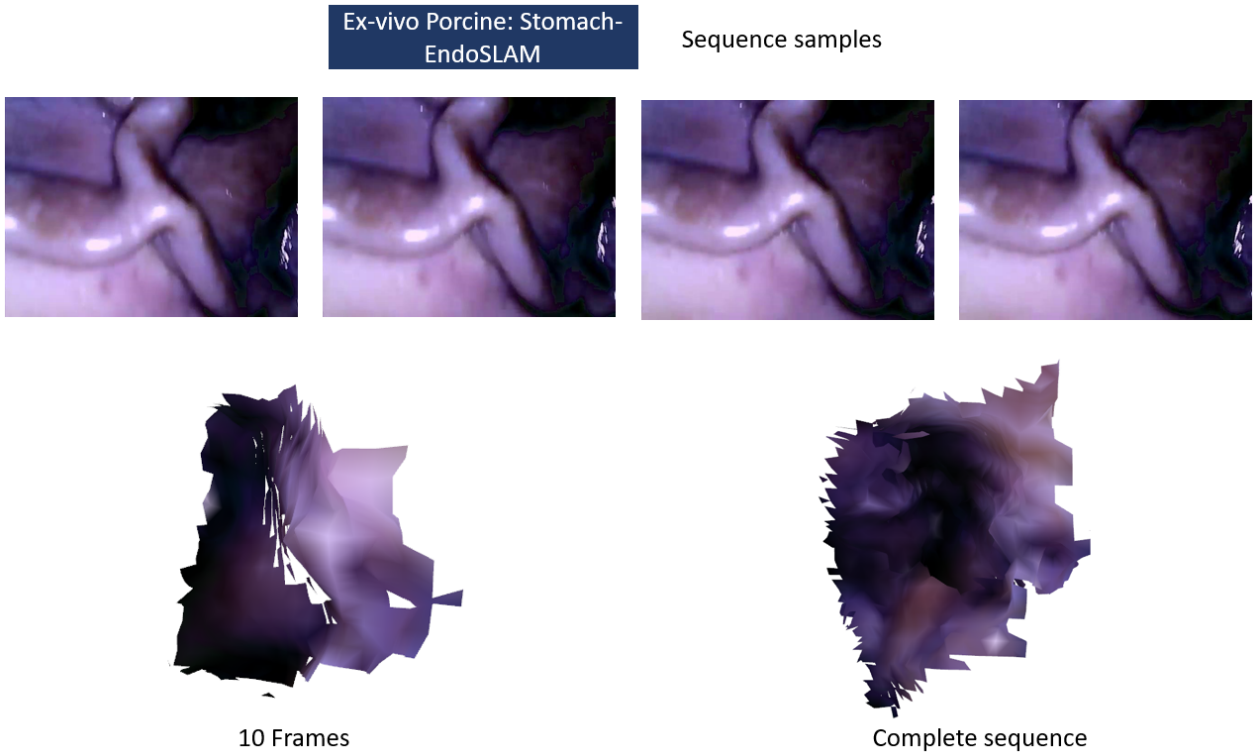


FIGURE 5.6 – Reconstruction results for a second ex-vivo porcine stomach sequence from the EndoSLAM dataset. Top: sequence samples showing mucosal folds and high surface reflectivity. Bottom: 3D reconstructions using 10 frames (left) and the complete sequence (right). The 10-frame result is sparse and fragmented due to specularities and limited overlap. The full-sequence reconstruction exhibits significantly improved coverage and shape continuity, although surface irregularities persist at the periphery.

motion blur and specular regions, but the overall geometry is topologically plausible.

These results highlight the effectiveness of the proposed pipeline in in-vivo conditions and the importance of temporal accumulation for structural completeness. They also demonstrate that the MonoIIT-based depth predictions maintain geometric consistency even under clinical variability, enabling reconstructions of entire colonic segments from monocular video.

Observations and Summary. Across all experiments, the addition of temporal information consistently improves reconstruction quality. The combination of MonoIIT's frame-wise depth accuracy and the Endo-Depth-and-Motion tracking framework enables the recovery of coherent 3D structures even under low-texture and narrow-baseline conditions. Notably, the pipeline exhibits good generalization across anatomical sites and acquisition modalities without retraining or domain-specific tuning.

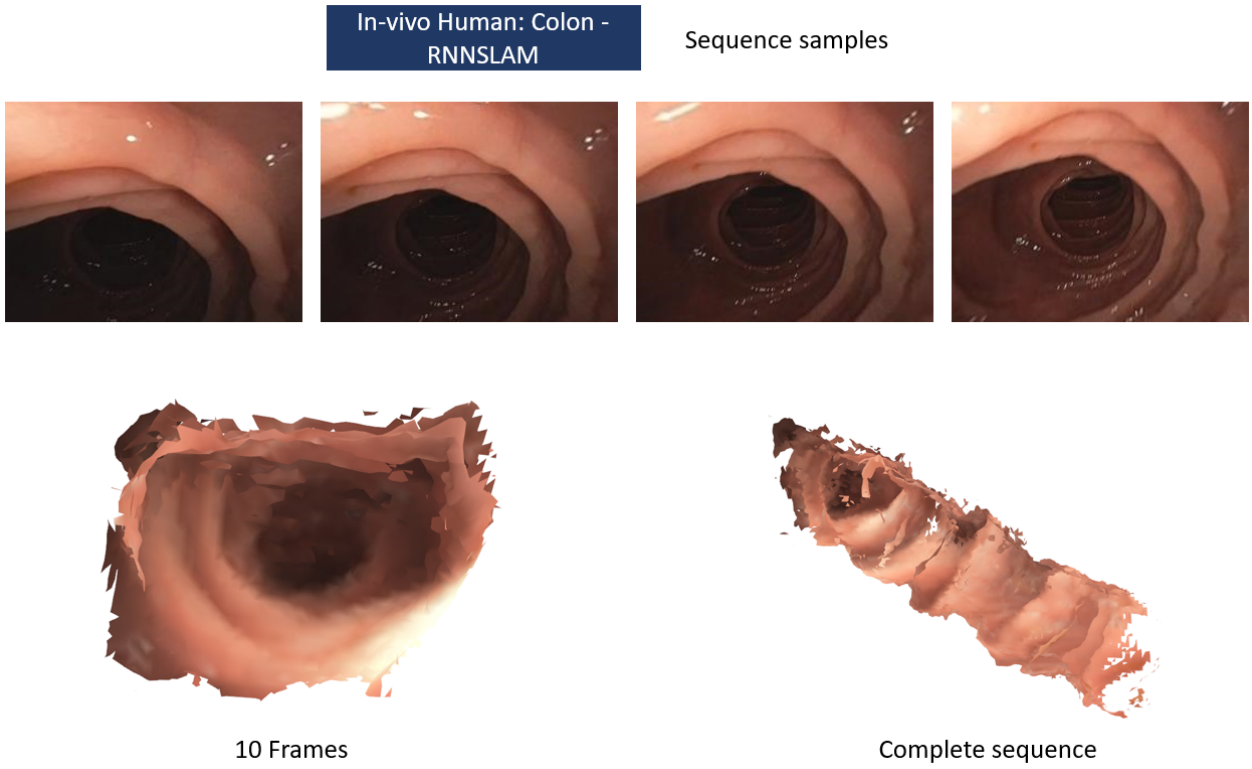


FIGURE 5.7 – Qualitative reconstruction results for an in-vivo human colon sequence from the RNNSLAM dataset. Top: representative RGB frames showing typical colonic folds and specular highlights. Bottom: 3D reconstructions using 10 frames (left) and the complete sequence (right). The full-sequence result captures a longer segment of the colon with improved geometric continuity and lumen structure, whereas the 10-frame reconstruction is more localized and exhibits minor surface discontinuities.

5.7 Discussion ◂

The qualitative experiments presented in the previous section demonstrate the feasibility of using a self-supervised monocular depth model, such as MonoIIT, in combination with photometric tracking and TSDF-based fusion, to reconstruct coherent 3D surfaces of different gastrointestinal organs. The approach shows notable adaptability across ex-vivo and in-vivo settings and across organs with diverse shapes and textures.

One of the key strengths observed is the generalization capability of MonoIIT across anatomical domains without additional fine-tuning. The predicted depth maps maintain surface continuity and preserve organ-specific geometric features, even in low-texture or poorly illuminated scenes. This is particularly evident in the in-vivo colon reconstructions, where depth predictions remained stable despite specularities, fluids, and strong deformations.

The integration of depth over time—moving from a single frame to a full sequence—markedly improves reconstruction quality. Temporal fusion mitigates the noise of single predictions and pro-

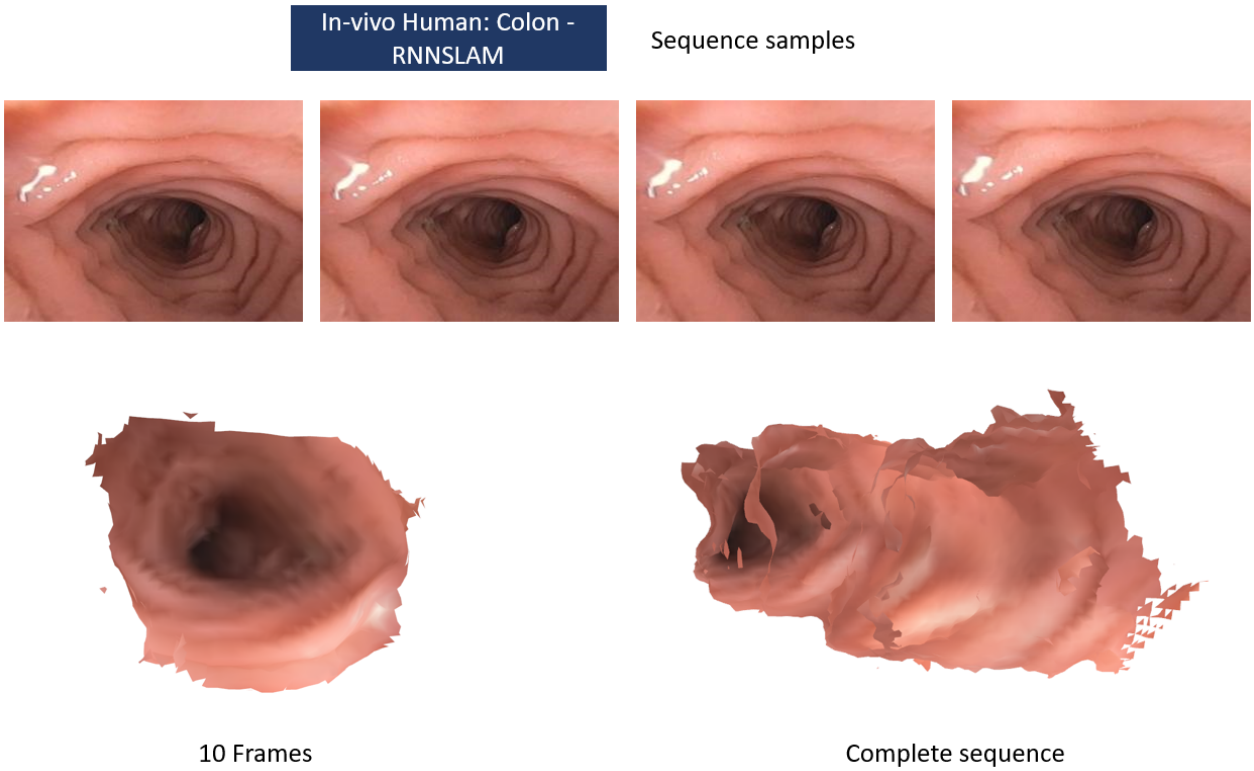


FIGURE 5.8 – Qualitative reconstruction results for a second in-vivo human colon sequence from the RNNSLAM dataset. Top: sample RGB frames showing colonic folds and specular highlights. Bottom: 3D reconstructions using 10 frames (left) and the complete sequence (right). The 10-frame reconstruction covers a limited region with incomplete geometry, while the full-sequence result recovers a longer colon segment with more consistent wall structure, despite residual tracking artifacts and illumination-induced distortion.

vides volumetric consistency. However, short sequences (e.g., 10 frames) are often insufficient for capturing the full lumen structure in highly folded or curved regions, especially in the stomach or colon. This highlights the need for sufficient spatial coverage and overlap to ensure reliable reconstruction.

Despite the robustness observed, several limitations remain. The reliance on photometric consistency assumes minimal lighting changes between frames, an assumption that does not always hold in in-vivo procedures. Furthermore, organ deformation, occlusions (e.g., from tools or fluids), and motion blur occasionally degrade pose estimation and fusion accuracy. While the proposed pipeline operates without explicit deformation modeling, incorporating such models could improve performance in highly dynamic regions.

Another challenge is the absence of ground truth for real-world datasets, which limits our evaluation to qualitative analysis. Future work could incorporate phantom-based or CT-registered datasets to enable a quantitative assessment of the reconstruction fidelity. Additionally, real-time implementation and integration into clinical workflows remain open directions.

In summary, the proposed pipeline offers a promising solution for monocular endoscopic 3D reconstruction across multiple organ systems. It leverages the strengths of self-supervised learning and photometric tracking while highlighting areas where domain-specific challenges call for further methodological improvements.

5.8 Conclusion ◂

In this chapter, we presented a 3D reconstruction pipeline for endoscopic imagery that integrates MonoIIT—a self-supervised depth prediction model—with a direct photometric tracking and volumetric fusion framework. The proposed system is designed to operate on monocular video and is tested across different gastrointestinal organs, including the small intestine and stomach (ex-vivo porcine), and the colon (in-vivo human).

Through several experiments, it has been demonstrated that the substitution of the original depth module in Endo-Depth-and-Motion with MonoIIT preserves the overall integrity of the pipeline, while enhancing its robustness to illumination variability and anatomical diversity. Through qualitative experiments on five representative sequences, it has been shown that even under challenging conditions, the proposed system is capable of generating consistent 3D reconstructions from monocular inputs.

These results highlight the potential of combining modern self-supervised models with SLAM-like pipelines for endoscopic navigation, scene understanding, and intra-operative decision support. The generalization to multiple organs and acquisition scenarios also suggests that such models could form the foundation of universal monocular 3D reconstruction systems in clinical endoscopy.

Conclusion and Perspectives

This thesis addresses the challenge of reconstructing extended, high-fidelity 3D surfaces from monocular endoscopic video sequences. The motivation arises from clinical needs for enhanced diagnostic accuracy, improved spatial interpretation of internal organ structures, and more effective post-procedural documentation in standard endoscopic practice. Traditional imaging workflows, which rely solely on 2D video navigation, provide limited spatial context and are prone to photometric inconsistencies that degrade the quality of downstream computer vision tasks such as frame registration, depth estimation, and 3D reconstruction.

Building upon recent advances in deep learning and geometric computer vision, this work proposes a novel passive vision pipeline that integrates photometric image enhancement and self-supervised depth estimation into a monocular reconstruction framework. Rather than relying on expensive or invasive depth-sensing hardware, the approach leverages neural models to generate photometrically consistent and geometrically plausible reconstructions from purely monocular video data. The proposed methods are trained and evaluated under both synthetic and clinically acquired conditions, with the aim of demonstrating their relevance for downstream 3D visualization tasks in gastroenterology.

The main contributions of this thesis can be summarized as follows:

— The creation of *Endo4IE*, a paired synthetic dataset specifically designed to support the training and evaluation of deep learning-based photometric image enhancement methods for endoscopy. The dataset includes realistic photometric distortions—such as underexposure, overexposure, and specular reflections—alongside clean ground truth counterparts, facilitating supervised training and quantitative benchmarking.

— The development of two learning-based enhancement models, *Endo-LMSPEC* and *Endo-ViT*, capable of restoring image quality and suppressing exposure-related artifacts and specular reflections in endoscopic video. These models are trained to enhance photometric consistency while preserving anatomical structures critical to downstream analysis.

— The integration of photometric enhancement modules into an incremental SLAM-based reconstruction pipeline, enabling improved surface reconstruction performance from monocular input. Although the depth estimation model was evaluated independently, the photometric corrections were shown to significantly improve tracking robustness and reconstruction quality

within the SLAM framework.

— The introduction of a self-supervised, illumination-invariant monocular depth estimation framework based on transformer architectures. The proposed model, tailored to endoscopic imagery, incorporates an illumination-invariant loss and local lighting calibration mechanism to improve depth prediction in the presence of texture sparsity and dynamic illumination conditions.

— A comprehensive experimental evaluation conducted on both synthetic and real-world endoscopic datasets, demonstrating the effectiveness of the proposed modules in improving image quality, reducing photometric noise, and enhancing the geometric accuracy of 3D reconstructions under realistic conditions.

Perspectives

Several research directions emerge from the findings of this thesis and may be explored in future work:

— *Scale Consistency in Monocular Depth Estimation:* One of the inherent limitations of monocular depth prediction is the ambiguity of global scale, particularly when depth models are trained in a self-supervised manner. While the current approach produces geometrically plausible reconstructions, scale drift across frames may impair long-term consistency and affect the accuracy of downstream tasks such as surface fusion or lesion localization. Future work could explore the integration of scale-aware constraints, such as geometric consistency losses over long sequences, depth normalization strategies, or weak supervision using endoscopic instruments of known size. Additionally, combining monocular depth cues with scale priors from SLAM trajectories or depth-from-focus signals may help resolve scale ambiguity in a temporally coherent manner.

— *Prompt-Assisted Illumination Artifact Correction:* Recent advances in clinician-in-the-loop enhancement systems, such as our prompt-assisted framework (Espinosa et al. [2025]), open new possibilities for adaptive pre-processing in endoscopic imaging. By leveraging a BERT-based model to interpret natural language prompts, the system dynamically selects and applies targeted correction techniques for overexposure, underexposure, and specular reflections. Unlike conventional global adjustments, this approach enables fine-grained, spatially localized corrections aligned with clinical intent, improving both visual interpretability and the robustness of downstream tasks such as SLAM-based 3D reconstruction and lesion detection. Future extensions could integrate this prompt-driven enhancement directly into real-time navigation pipelines, allowing interactive refinement of image quality during live procedures.

— *Deformation modeling:* Although the current framework assumes quasi-rigid organ surfaces, real colonoscopic sequences often exhibit dynamic deformations due to peristalsis, patient

breathing, or simply the intrinsic softness of the tissue. A promising direction would be the integration of non-rigid SLAM or dynamic scene reconstruction techniques, such as deformation graphs or locally rigid surface models, to capture temporal and spatial variations in tissue geometry. Physics-inspired priors or biomechanical constraints could also be incorporated to improve realism and convergence stability.

— *Uncertainty estimation:* Incorporating confidence modeling into both depth and pose estimation could increase robustness, especially under degraded visual conditions. Probabilistic depth networks, such as those based on Bayesian deep learning or dropout-based Monte Carlo sampling, may be explored. These confidence maps could be propagated during 3D fusion to adaptively weight unreliable regions, suppressing noise and improving mesh consistency.

— *Cross-modal registration:* Particularly in the context of laparoscopic or multimodal imaging, the reconstructed 3D surfaces could be registered to pre-operative CT or MRI volumes. Scientific avenues here include the use of learned feature descriptors for multi-modal correspondence, and surface-to-volume registration using differentiable renderers. Such alignment would enable hybrid navigation systems and spatially grounded semantic interpretation for surgical guidance.

— *Clinical validation and deployment:* The clinical potential of the proposed pipeline warrants extensive prospective validation. One concrete direction is the development of a semi-automated lesion mapping tool based on the 3D reconstructions, to be tested in a longitudinal cohort. This would involve defining spatial reproducibility metrics, usability assessments with gastroenterologists, and measuring clinical impact on lesion detection and documentation during follow-up procedures.

In conclusion, the methodology developed in this thesis lays the foundation for photometrically and geometrically consistent 3D reconstruction in colonoscopy. It bridges the gap between image enhancement and spatial modeling, offering a promising step toward intelligent, vision-driven computer-assisted endoscopy systems.

List of Publications ◂

International journals

— **Transformer-Based Illumination Invariant Self-Supervised Monocular Depth Estimation in Endoscopy**
Ricardo Espinosa, Gilberto Ochoa-Ruiz, Christian Daul
Preprint sent to CVIU (Computer Vision and Image Understanding).

International conferences

1. **A Novel Hybrid Endoscopic Dataset for Evaluating Machine Learning-Based Photometric Image Enhancement Models**
Carlos Axel Garcia-Vega, Ricardo Espinosa, Gilberto Ochoa-Ruiz, Thomas Bazin, Luis Eduardo Falcón-Morales, Dominique Lamarque, Christian Daul
Advances in Computational Intelligence – MICAI 2022, Lecture Notes in Computer Science, vol. 13612, Springer, 2022, pp. 267–281.
2. **Multi-Scale Structural-Aware Exposure Correction for Endoscopic Imaging**
A. Garcia-Vega, R. Espinosa, L. Ramirez-Guzman, T. Bazin, L. Falcón-Morales, G. Ochoa-Ruiz, D. Lamarque, and C. Daul.
In: *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, 2023.
3. **Color-Aware Exposure Correction for Endoscopic Imaging Using a Lightweight Vision Transformer**
Ricardo Espinosa, Javier Eluney Hernández, Gilberto Ochoa-Ruiz, Christian Daul
IEEE 37th International Symposium on Computer-Based Medical Systems (CBMS), 2024, pp. 376–382.
4. **A Deep Learning-Based Image Pre-Processing Pipeline for Enhanced 3D Colon Surface Reconstruction Robust to Endoscopic Illumination Artifacts**, Ricardo Espinosa, Carlos Axel García-Vega, Gilberto Ochoa-Ruiz, Dominique Lamarque, Christian Daul.
In: *Proceedings of the 2024 IEEE 37th International Symposium on Computer-Based Medical Systems (CBMS)*. https://ieeexplore.ieee.org/document/10234568
5. **EndoDepth: A Benchmark for Assessing Robustness in Endoscopic Depth Prediction**
Ivan Reyes-Amezcua, Ricardo Espinosa, Christian Daul, Gilberto Ochoa-Ruiz, Andres Mendez-Vazquez
In: *MICCAI Workshop on Data Engineering in Medical Imaging (DEMI 2024)*, Lecture Notes in Computer Science, Springer, 2024, pp. 84–94. First Online: 25 October 2024.
6. **Prompt Assisted Enhancement for Correcting Illumination Artifacts in Endoscopic Images**
Ricardo Espinosa, Gilberto Ochoa-Ruiz, Christian Daul
Accepted for oral presentation at MICAI 2025. To be published in the Springer LNAI series.

National conferences

1. **Deep Learning-Based Image Exposure Enhancement as a Pre-Processing for an Accurate 3D Colon Surface Reconstruction**,
   Ricardo Espinosa, Carlos Axel García-Vega, Gilberto Ochoa-Ruiz, Dominique Lamarque, Christian Daul.
   In: *Proceedings of the 2023 GRETSI Symposium on Signal and Image Processing*

Bibliography

Afifi, M. and Brown, M. S. (2021). Histogan: Controlling colors of gan-generated and real images via color histograms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7949–7958.

Afifi, M., Derpanis, K. G., Ommer, B., and Brown, M. S. (2021). Learning multi-scale photo exposure correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9157–9167.

Agarwal, S., Snavely, N., Simon, I., Seitz, S. M., and Szeliski, R. (2009). Building rome in a day. In *2009 IEEE 12th International Conference on Computer Vision*, pages 72–79.

Ali, S. (2016). *Total variational optical flow for robust and accurate bladder image mosaicing*. Ph.D. thesis, Université de Lorraine, France.

Ali, S., Daul, C., Galbrun, E., and Blondel, W. (2016a). Illumination invariant optical flow using neighborhood descriptors. *Computer Vision and Image Understanding*, 145:95–110.

Ali, S., Daul, C., Galbrun, E., Guillemin, F., and Blondel, W. (2016b). Anisotropic motion estimation on edge preserving riesz wavelets for robust video mosaicing. *Pattern Recognition*, 51:425–442.

Ali, S., Dmitrieva, M., Zhou, F., Daul, C., Braden, B., Bailey, A., East, J., Realdon, S., Wagnieres, G., Loshchenov, M., Blondel, W., Grisan, E., and Rittscher, J. (2021a). Endoscopy artifact detection (ead) dataset (includes updated 2020 version). Version 4.

Ali, S., Zhou, F., Bailey, A., Braden, B., East, J., Lu, X., and Rittscher, J. (2021b). A deep learning framework for quality assessment and restoration in video endoscopy. *Medical Image Analysis*, 68:101900.

Ali, S., Zhou, F., Braden, B., Bailey, A., Yang, S., Cheng, G., Pengyi Zhang, X. L., Kayser, M., Soberanis-Mukul, R. D., Albarqouni, S., Wang, X., Wang, C., Watanabe, S., Oksuz, I., Ning, Q., Yang, S., Khan, M. A., Gao, X. W., Realdon, S., Maxim Loshchenov, J. A. S., East, J. E., Wagnieres, G., Victor B. Loschenov, E. G., Daul, C., Blondel, W., and Rittscher, J. (2020). An

objective comparison of detection and segmentation algorithms for artifacts in clinical endoscopy. *Scientific Reports*, 10:2748.

Allan, M., McLeod, A. J., Wang, C., Rosenthal, J., Hu, Z., Gard, N., Eisert, P., Fu, K. X., Zeffiro, T., Xia, W., Zhu, Z., Luo, H., Jia, F., Zhang, X., Li, X., Sharan, L., Kurmann, T., Schmid, S., Sznitman, R., Psychogyios, D., Azizian, M., Stoyanov, D., Maier-Hein, L., and Speidel, S. (2021). Stereo correspondence and reconstruction of endoscopic data challenge. *CoRR*, abs/2101.01133.

Almalioglu, Y., Ozyoruk, K. B., Gokce, A., Incetan, K., Gokceler, G. I., Simsek, M. A., Ararat, K., Chen, R. J., Durr, N. J., Mahmood, F., and Turan, M. (2020a). Endol2h: Deep super-resolution for capsule endoscopy. *Medical Image Analysis*, 64:101718.

Almalioglu, Y., Turan, M., Gilbert, H., Patané, G., Navab, N., and Turan, M. (2020b). Endol2h: Robust image enhancement for real-time endoscopic navigation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 279–289. Springer.

Anoop, P. and Deivanathan, R. (2024). Advancements in low light image enhancement techniques and recent applications. *Journal of Visual Communication and Image Representation*, 103:104223.

Bartoli, A., Gérard, Y., Chadebecq, F., and Collins, T. (2012). On template-based reconstruction from a single view: Analytical solutions and proofs of well-posedness for developable, isometric and conformal surfaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2026–2033. IEEE Computer Society.

Behrens, A., Stehle, T., Gross, S., and Aach, T. (2009). Local and global panoramic imaging for fluorescence bladder endoscopy. In *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 6990–6993.

Ben-Hamadou, A., Daul, C., and Soussen, C. (2016). Construction of extended 3D field of views of the internal bladder wall surface: A proof of concept. *3D Research*, 7(3):95:1–95:23.

Ben-Hamadou, A., Soussen, C., Daul, C., Blondel, W., and Wolf, D. (2013). Flexible calibration of structured-light systems projecting point patterns. *Computer Vision and Image Understanding*, 117(10):1468–1481.

Bernal, J., Tajkbaksh, N., Sánchez, F., Matuszewski, B., Chen, H., Yu, L., Angermann, Q., Romain, O., Rustad, R., Balasingham, I., et al. (2018). Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics*, 69:77–87.

Blau, Y. and Michaeli, T. (2018). Perception-distortion tradeoff in image super-resolution. In *ECCV Workshops*, pages 1–17. Springer.

Bonilla, S., Zhang, S., Psychogyios, D., Stoyanov, D., Vasconcelos, F., and Bano, S. (2024). Gaussian pancakes: Geometrically-regularized 3D gaussian splatting for realistic endoscopic reconstruction. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume LNCS 15006, pages 274–283. Springer Nature Switzerland.

Borgli, H. et al. (2020). Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific Data*, 7:283.

Bregler, C., Hertzmann, A., and Biermann, H. (2000). Recovering non-rigid 3D shape from image streams. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2690–2696. IEEE.

Bruckstein, S. and Netravali, A. N. (1992). Shape from shading using level curves. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 302–307.

Bychkovsky, V., Paris, S., Chan, E., and Durand, F. (2011). Learning photographic global tonal adjustment with a database of input/output image pairs. *ACM Transactions on Graphics (TOG)*, 30(2):1–12.

Cappell, M. S. (2008). The role of endoscopy in the diagnosis and management of gastrointestinal disorders. *The New England Journal of Medicine*, 359(20):2128–2139.

Carroll, R. E. and Seitz, S. M. (2007). Rectified surface mosaics. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8.

Chen, C., Chen, Q., Xu, J., and Koltun, V. (2018a). Learning to see in the dark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3291–3300.

Chen, Y., Zhu, Y., Zhang, S., Metaxas, D. N., and Yang, X. (2018b). Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans. In *European Conference on Computer Vision (ECCV)*, pages 289–305. Springer.

Cheng, J. and Alkaisi, A. (2012). Real-time 3D surface recovery in laparoscopic video using SfS. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*.

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

Cohen, J. P., Luck, M., and Honari, S. (2018). Distribution matching losses can hallucinate features in medical image translation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 529–536. Springer.

Cryer, J. E., Tsai, P. S., and Shah, M. (1995). Integration of shape from shading and stereo. *Pattern Recognition*, 28(7):1033–1043.

Cui, Z. et al. (2022). You only need 90k parameters to adapt light: A lightweight transformer for image enhancement and exposure correction. In *BMVC*.

Daher, R., Vasconcelos, F., and Stoyanov, D. (2023). A temporal learning approach to inpainting endoscopic specularities and its effect on image correspondence. *Medical Image Analysis*, 90:102994.

Daul, C., Graebling, P., Tiedeu, A., and Wolf, D. (2005). 3-D reconstruction of microcalcification clusters using stereo imaging: algorithm and mammographic unit calibration. *IEEE Transactions on Biomedical Engineering*, 52(12):2058–2073.

Davison, A. J. (2003). Real-time simultaneous localisation and mapping with a single camera. In *Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV)*, pages 1403–1410. IEEE Computer Society.

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

Diamantas, S. C., Oikonomidis, A., and Crowder, R. M. (2010). Depth estimation for autonomous robot navigation: A comparative approach. In *2010 IEEE International Conference on Imaging Systems and Techniques*, pages 426–430.

Dietrich, J. T. (2016). Riverscape mapping with helicopter-based structure-from-motion photogrammetry. *Geomorphology*, 252:144–157.

Doutre, C. and Nasiopoulos, P. (2009). Fast vignetting correction and color matching for panoramic image stitching. In *16th IEEE International Conference on Image Processing (ICIP)*, pages 709–712.

Edwards, P. E., Psychogyios, D., Speidel, S., Maier-Hein, L., and Stoyanov, D. (2022). Serv-ct: A disparity dataset from cone-beam ct for validation of endoscopic 3d reconstruction. *Medical Image Analysis*, 76:102302.

Espinosa, R., Cerriteño, J., González-Domínguez, S., Ochoa-Ruiz, G., and Daul, C. (2024a). A deep learning-based image pre-processing pipeline for enhanced 3D colon surface reconstruction robust to endoscopic illumination artifacts. In *Proceedings of the 37th IEEE International Symposium on Computer-Based Medical Systems (CBMS)*, pages 81–88.

Espinosa, R., García-Vega, C. A., Ochoa-Ruiz, G., Lamarque, D., and Daul, C. (2023). Deep learning-based image exposure enhancement as a pre-processing for an accurate 3D colon surface reconstruction. In *GRETSI Symposium on Signal and Image Processing*.

Espinosa, R., Hernandez, E., Ochoa-Ruiz, G., and Daul, C. (2024b). Color-aware exposure correction for endoscopic imaging using a lightweight vision trans- former. In *Proceedings of the 37th IEEE International Symposium on Computer-Based Medical Systems (CBMS)*, pages 376–382.

Espinosa, R., Hernández, E., Magaña, J. C., Ochoa, G., and Daul, C. (2025). Prompt assisted enhancement for correcting illumination artifacts in endoscopic images. In *Proceedings of the 24th Mexican International Conference on Artificial Intelligence (MICAI 2025)*, volume 16221 of *Lecture Notes in Artificial Intelligence (LNAI)*, pages 40–54. Springer.

Foix, S., Alenyà, G., and Torras, C. (2011). Lock-in time-of-flight (ToF) cameras: A survey. *IEEE Sensors Journal*, 11(9):1917–1926.

Forster, M. and Thiran, P. (2000). Endoscopic surface reconstruction using a dichromatic model and shape-from-shading. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*.

Fu, Y., Liu, S., Kulkarni, A., Kautz, J., Efros, A. A., and Wang, X. (2023). Colmap-free 3d gaussian splatting.

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

Graebling, P., Boucher, C., Daul, C., and Hirsch, E. (1995). 3D sculptured surface analysis using a structured-light approach. In *Videometrics IV*, volume 2598, pages 128–139. SPIE.

Grasa, O. G., Bernal, E., Casado, S., Gil, I., and Montiel, J. M. M. (2014). Visual SLAM for handheld monocular endoscope. *IEEE Transactions on Medical Imaging*, 33(1):135–146.

Guo, C., Li, C., Guo, J., Loy, C. C., Hou, J., Kwong, S., and Cong, R. (2020). Zero-reference deep curve estimation for low-light image enhancement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1780–1789. IEEE.

Guo, X., Li, Y., and Ling, H. (2017a). Lime: Low-light image enhancement via illumination map estimation. *IEEE Transactions on Image Processing*, 26(2):982–993.

Guo, X., Li, Y., and Ling, H. (2017b). Lime: Low-light image enhancement via illumination map estimation. *IEEE Transactions on Image Processing*, 26(2):982–993.

Haase, S. and Maier, A. (2018). Endoscopy. In Maier, A., Steidl, S., Christlein, V., and Hornegger, J., editors, *Medical Imaging Systems*, volume 11111 of *Lecture Notes in Computer Science*, pages 57–68. Springer.

Hansard, M., Lee, S., Choi, O., and Horaud, R. (2013). *Time-of-Flight Cameras: Principles, Methods and Applications*. SpringerBriefs in Computer Science. Springer.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Heinly, J., Schönberger, J. L., Dunn, E., and Frahm, J.-M. (2015). Reconstructing the world* in six days *(as captured by the yahoo 100 million image dataset). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3287–3295.

Horn, B. K. P. (1975). Obtaining shape from shading information. *The Psychology of Computer Vision*, pages 115–155.

Ignatov, A., Kobyshev, N., Timofte, R., Vanhoey, K., and Van Gool, L. (2017). Dslr-quality photos on mobile devices with deep convolutional networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3277–3285. IEEE.

James, M. R. and Robson, S. (2012). Straightforward reconstruction of 3D surfaces and topography with a camera: Accuracy and geoscience application. *Journal of Geophysical Research: Earth Surface*, 117:F03017:1–17.

Jawdekar, A. S. and Joshi, A. P. (2021). A review on image enhancement techniques. *Materials Today: Proceedings*, 47:4666–4672.

Jiang, D., Zhang, J., Qin, H., Wang, X., Wang, B., and Jiang, T. (2021a). A deep learning approach for the classification of gastrointestinal diseases using wireless capsule endoscopy images. In *Computers in Biology and Medicine*, volume 133, page 104399. Elsevier.

Jiang, Y., Gong, X., Liu, D., Cheng, Y., Fang, C., Shen, X., Zhou, P., Fang, C., and Yang, Z. (2021b). Enlightengan: Deep light enhancement without paired supervision. In *British Machine Vision Conference (BMVC)*.

Kohler, T., Scherl, H., and Hornegger, J. (2014). Mask-specific inpainting and denoising of endoscopic images. *IEEE Transactions on Image Processing*, 23(7):2961–2973.

Kuo, B. W., Chang, H. H., Chen, Y. C., and Huang, S. Y. (2011). A light-and-fast SLAM algorithm for robots in indoor environments using line segment map. *Journal of Robotics*, 2011:257852:1–257852:12.

Land, E. H. and McCann, J. J. (1971). Lightness and retinex theory. In *Journal of the Optical Society of America*, volume 61, pages 1–11. Optical Society of America.

Lee, T. and Rosenfeld, S. (1985). Improved methods of shape from shading using the light source coordinate system. *Artificial Intelligence*, 26(2):125–143.

Lhuillier, M. and Quan, L. (2000). Robust dense matching using local and global geometric constraints. In *Proceedings of the 15th International Conference on Pattern Recognition (ICPR)*, pages 1968–1972. IEEE Computer Society.

Li, L., Li, X., Yang, S., Ding, S., Jolfaei, A., and Zheng, X. (2021). Unsupervised learning-based continuous depth and motion estimation with monocular endoscopy for virtual reality minimally invasive surgery. *IEEE Transactions on Industrial Informatics*, 17(6):3920–3928.

Lindner, M., Schiller, I., Kolb, A., and Koch, R. (2010). Time-of-flight sensor calibration for accurate range sensing. *Computer Vision and Image Understanding*, 114(12):1318–1328.

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

Liu, Y., Li, C., Yang, C., and Yuan, Y. (2024). Endogaussian: Real-time gaussian splatting for dynamic endoscopic scene reconstruction. *arXiv preprint*, arXiv:2401.12561.

Lorensen, W. E. and Cline, H. E. (1987). Marching cubes: A high resolution 3d surface construction algorithm. In Stone, M. C., editor, *SIGGRAPH*, pages 163–169. ACM.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110.

Luo, H., Gao, Y., Wu, Y., Liao, C., Yang, X., and Cheng, K.-T. (2019). Real-time dense monocular slam with online adapted depth prediction network. *IEEE Transactions on Multimedia*, 21(2):470–483.

Lurie, K., Al-Naji, A., Zhang, Y., and Han, J. (2017a). Dense 3D reconstruction of bladder wall surface from clinical cystoscopic videos using structure-from-motion. In *Proceedings of the IEEE Engineering in Medicine and Biology Conference (EMBC)*, pages 1284–1287. IEEE.

Lurie, Y., Smith, A., and Johnson, B. (2017b). Advanced 3D reconstruction techniques in endoscopic imaging. *Journal of Medical Imaging*, 24(3):123–130.

Lv, F., Lu, F., Wu, J., Lim, S., and Yu, H. (2020). Fast and accurate image enhancement with learnable exposure adjustment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10075–10084. IEEE.

Ma, C., Yang, C.-Y., Yang, X., and Yang, M.-H. (2017). Learning a no-reference quality metric for single-image super-resolution. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4945–4953. IEEE.

Ma, J., Ali, S., Ciaramella, A., Gurudu, S., and Rittscher, J. (2022). Self-supervised learning for dense depth estimation in monocular endoscopy. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 294–304. Springer.

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

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

Mahmood, F., Chen, H., Durr, N., Raza, S., et al. (2022). Deep learning-based 3d reconstruction for colonoscopy: a review. *IEEE Transactions on Medical Imaging*, 41(1):1–19.

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

Maier-Hein, L., Mountney, P., Bartoli, A., Elhawary, H., Elson, D. S., Groch, A., Kolb, A., Rodrigues, M. A., Sorger, J. M., Speidel, S., and Stoyanov, D. (2013). Optical techniques for 3D surface reconstruction in computer-assisted laparoscopic surgery. *Medical Image Analysis*, 17(8):974–996.

Martínez, A., Trinh, D.-H., El Beze, J., Hubert, J., Eschwege, P., Estrade, V., Aguilar, L., Daul, C., and Ochoa, G. (2020). Towards an automated classification method for ureteroscopic kidney

stone images using ensemble learning. In *International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1936–1939.

Miao, X. Y., Miao, X. N., Ye, L. Y., and Cheng, H. (2021). Image enhancement model based on deep learning applied to the ureteroscopic diagnosis of ureteral stones during pregnancy. *Computational and Mathematical Methods in Medicine*, 2021:9548312.

Miranda-Luna, R., Daul, C., Blondel, W. C. P. M., Hernandez-Mier, Y., Wolf, D., and Guillemin, F. (2008). Mosaicing of bladder endoscopic image sequences: Distortion calibration and registration algorithm. *IEEE Transactions on Biomedical Engineering*, 55(2):541–553.

Miranda-Luna, R., Hernandez-Mier, Y., Daul, C., Blondel, W., and Wolf, D. (2004). Mosaicing of medical video-endoscopic images: data quality improvement and algorithm testing. In *(ICEEE). 1st International Conference on Electrical and Electronics Engineering, 2004.*, pages 530–535.

Mittal, A., Moorthy, A. K., and Bovik, A. C. (2012). Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212.

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

Pahlavan, A., Forster, M., Fong, N., and Kim, J. (2012). 3D reconstruction in capsule endoscopy using shape-from-shading. In *Proceedings of the IEEE Engineering in Medicine and Biology Conference (EMBC)*.

Penedo, T. S. (1988). Shape from shading using linear approximation. *Image and Vision Computing*, 6(3):143–150.

Penne, J., Höller, K., Stürmer, M., Schrauder, T., Schneider, A., Engelbrecht, R., Feußner, H., Schmauss, B., and Hornegger, J. (2009). Time-of-Flight 3-D endoscopy. In Yang, G.-Z., Hawkes, D., Rueckert, D., Noble, A., and Taylor, C., editors, *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 467–474, Berlin, Heidelberg. Springer Berlin Heidelberg.

Phan, T.-B. (2020). *On the 3D hollow organ cartography using 2D endoscopic images*. Ph.D. thesis, Université de Lorraine. Image Processing [eess.IV].

Phan, T.-B., Trinh, D.-H., Lamarque, D., Wolf, D., and Daul, C. (2019). Dense optical flow for the reconstruction of weakly textured and structured surfaces: Application to endoscopy. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 310–314.

Phan, T.-B., Trinh, D.-H., Wolf, D., and Daul, C. (2020). Optical flow-based structure-from-motion for the reconstruction of epithelial surfaces. *Pattern Recognition*, 105:107391.

Qin, Y., Wang, Y., Zhang, Z., and Zhang, M. (2024). Endoscope-nerf: Neural radiance fields for reconstructing deformable tissues in endoscopic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12345–12354. IEEE.

Recasens, A., Butterfield, Y., Hager, G. D., and Unberath, M. (2021). Endo-depth: Real-time depth estimation from monocular endoscopy. *IEEE Robotics and Automation Letters*, 6(2):2921–2928.

Reyes-Amezcua, I., Espinosa, R., Daul, C., Ochoa-Ruiz, G., and Mendez-Vazquez, A. (2024). Endodepth: A benchmark for assessing robustness in endoscopic depth prediction. In *Data Engineering in Medical Imaging (DEMI 2024) - MICCAI Workshop*, Lecture Notes in Computer Science, pages 84–94. Springer. First Online: 25 October 2024.

Röhl, S., Bodenstedt, S., Suwelack, S., Kenngott, H., Müller-Stich, B. P., Dillmann, R., and Speidel, S. (2012). Dense gpu-enhanced surface reconstruction from stereo endoscopic images for intraoperative registration. *Medical Physics*, 39(3):1632–1645.

Rossi, R., Savatier, X., Ertaud, J. Y., and Mazari, B. (2009). Real-time 3D reconstruction for mobile robot using catadioptric cameras. In *Proceedings of the IEEE International Symposium on Robotic and Sensors Environments (ROSE)*, pages 104–109. IEEE.

Scharstein, D. and Szeliski, R. (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1–3):7–42.

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

Seshamani, S., Lau, W., and Hager, G. (2006). Real-time endoscopic mosaicking. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 355–363. Springer, Berlin, Heidelberg.

Shao, S., Pei, Z., Chen, W., Zhu, W., Wu, X., Sun, D., and Zhang, B. (2022). Self-supervised monocular depth and ego-motion estimation in endoscopy: appearance flow to the rescue. *Medical Image Analysis*, 77.

Sharafutdinov, D., Griguletskii, M., Kopanev, P., Kurenkov, M., Ferrer, G., Burkov, A., Gonnochenko, A., and Tsetserukou, D. (2023). Comparison of modern open-source visual slam approaches.

Shevchenko, N., Fallert, J., Stepp, H., Sahli, H., Karl, A., and Lueth, T. (2012). A high resolution bladder wall map: Feasibility study. In *34th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5761–5764, San Diego, CA, USA.

Soper, T., Lee, C., and Kim, D. (2012a). Real-time 3D surface reconstruction in endoscopic procedures. In *Proceedings of the International Conference on Medical Imaging*, pages 456–462. Medical Imaging Society.

Soper, T., Porter, M., and Seibel, E. (2012b). Surface mosaics of the bladder reconstructed from endoscopic video for automated surveillance. *IEEE Transactions on Biomedical Engineering*, 59(6):1670–1680.

Stoyanov, D. (2012). Stereoscopic scene flow for robotic assisted minimally invasive surgery. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 7510 of *Lecture Notes in Computer Science*, pages 479–486. Springer.

Stoyanov, D., Scarzanella, M. V., Pratt, P., and Yang, G. Z. (2010). Real-time stereo reconstruction in robotically assisted minimally invasive surgery. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 6361 of *Lecture Notes in Computer Science*, pages 275–282. Springer.

Tao, X., Gao, H., Shen, X., Wang, J., and Jia, J. (2018). Scale-recurrent network for deep image deblurring. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8174–8182. IEEE.

Trinh, D. H. and Daul, C. (2019). On illumination-invariant variational optical flow for weakly textured scenes. *Computer Vision and Image Understanding*, 179:1–18.

Trinh, D. H., Daul, C., Blondel, W., and Lamarque, D. (2018). Mosaicing of images with few textures and strong illumination changes: Application to gastroscopic scenes. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 1263–1267.

Vega, C. A. G., Espinosa, R., Ochoa-Ruiz, G., Bazin, T., Morales, L. E. F., Lamarque, D., and Daul, C. (2022). A novel hybrid endoscopic dataset for evaluating machine learning-based photometric image enhancement models. In *Proceedings of the 21st Mexican International Conference on Artificial Intelligence (MICAI)*, Lecture Notes in Artificial Intelligence. Springer.

Vega, C. A. G., Espinosa, R., Ochoa-Ruiz, G., Bazin, T., Morales, L. E. F., Lamarque, D., and Daul, C. (2023). Multi-scale structural-aware exposure correction for endoscopic imaging. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5.

Viola, P. and Wells, W. (1995). Alignment by maximization of mutual information. In *Proceedings of IEEE International Conference on Computer Vision*, pages 16–23.

Wang, C., Peng, J., and Li, Z. (2008). Flattest histogram specification with accurate brightness preservation. *IET Image Processing*, 2(5):249–262.

Wang, F., Wang, J., and Xia, Y. (2022a). Illumination correction in endoscopic images using generative adversarial networks. *Medical Image Analysis*, 77:102362.

Wang, L., Niu, Y., and Jiang, Y. (2010). Shape from shading techniques for bone surface reconstruction in endoscopy. *Medical Physics*.

Wang, R., Pizer, S. M., and Frahm, J.-M. (2019). Recurrent neural network for (un-)supervised learning of monocular video visual odometry and depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612.

Wei, C., Wang, W., Yang, W., and Liu, J. (2018). Deep retinex decomposition for low-light enhancement. In *British Machine Vision Conference (BMVC)*.

Weibel, T., Daul, C., Wolf, D., and Rösch, R. (2012a). Contrast-enhancing seam detection and blending using graph cuts. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR)*, pages 2732–2735.

Weibel, T., Daul, C., Wolf, D., Rösch, R., and Guillemin, F. (2012b). Graph based construction of textured large field of view mosaics for bladder cancer diagnosis. *Pattern Recognition*, 45(12):4138–4150.

Widya, A. R., Monno, Y., Okutomi, M., Suzuki, S., Gotoda, T., and Miki, K. (2019). Whole stomach 3D reconstruction and frame localization from monocular endoscope video. *IEEE Journal of Translational Engineering in Health and Medicine*, 7:1–10.

Widya, A. R., Monno, Y., Okutomi, M., Suzuki, S., Gotoda, T., and Miki, K. (2021). Stomach 3D reconstruction using virtual chromoendoscopic images. *IEEE Journal of Translational Engineering in Health and Medicine*, 9:1–11.

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

Yang, S.-P., Kwon, J.-M., Seo, J.-W., Jun, H. J., Hwang, K., Kim, E.-S., and Jeong, K.-H. (2023). Light switching microprojector allows endoscopic in vivo 3d imaging of gastrointestinal abnormalities. *Advanced Photonics Research*, 4(2):2200254.

Yeung, D., Tsoi, H., and Yung, K. (1999). Shape-from-shading reconstruction of endoscopic surfaces using singular point detection. *Pattern Recognition Letters*, 20(5):463–472.

Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., and Darrell, T. (2020). Bdd100k: A diverse driving video database with scalable annotation tooling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2636–2645.

Zanchi, A., Salvi, F., Zanchetta, S., Sterlacchini, S., and Guerra, G. (2009). 3D reconstruction of complex geological bodies: Examples from the alps. *Computers & Geosciences*, 35(1):49–69.

Zenteno, O., Trinh, D.-H., Treuillet, S., Lucas, Y., Bazin, T., Lamarque, D., and Daul, C. (2022). Optical biopsy mapping on endoscopic image mosaics with a marker-free probe. *Computers in Biology and Medicine*, 143:105234.

Zhang, Q., Nie, Y., and Zheng, W.-S. (2019a). Dual illumination estimation for robust exposure correction. *Computer Graphics Forum*, 38(1):243–252.

Zhang, R. and Cryer, J. E. (1991). Surface shape from shading using energy minimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 534–539.

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

Zhang, Y., Wang, S., Ma, R., McGill, S. K., Rosenman, J. G., and Pizer, S. M. (2021). Lighting enhancement aids reconstruction of colonoscopic surfaces. In *Information Processing in Medical Imaging (IPMI)*, volume 12729 of *Lecture Notes in Computer Science*, pages 559–570. Springer.

Zhang, Y., Zhang, Y., Timofte, R., and Van Gool, L. (2019b). Kindling the darkness: A practical low-light image enhancer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2281–2290. IEEE.

Zhao, C., Zhang, Y., Poggi, M., Tosi, F., Guo, X., Zhu, Z., Huang, G., Tang, Y., and Mattoccia, S. (2022). Monovit: Self-supervised monocular depth estimation with a vision transformer. In *2022 International Conference on 3D Vision (3DV)*. IEEE.

Zhao, R., Wang, Y., Zheng, J., Song, Q., Wang, X., and Guo, G. (2021). Automated tissue classification in colonoscopy using deep learning and 3D conditional random fields. *IEEE Transactions on Medical Imaging*, 40(5):1325–1335.

Zhong, Z., Han, H., Zheng, Y., Wei, C., Wang, W., and Liu, J. (2020). Kind: Towards high-quality low-light image enhancement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9444–9452. IEEE.

Zhou, J., Wang, Y., Qin, K., and Zeng, W. (2019). Moving indoor: Unsupervised video depth learning in challenging environments. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8617–8626.

Zhou, L., Wu, G., Zuo, Y., Chen, X., and Hu, H. (2024). A comprehensive review of vision-based 3D reconstruction methods. *Sensors*, 24(7):2314.

Zhou, Q.-Y., Park, J., and Koltun, V. (2018). Open3d: A modern library for 3d data processing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages Open3D:1–8.

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

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2223–2232.

Zuo, C., Chen, Q., and Sun, S. (2013). Range-limited bi-histogram equalization for image contrast enhancement. *Optik*, 124(5):425–431.