Deep Learning Approaches for the Localization of Capsule Endoscope

by

Kutsev Bengisu Özyörük

B.S., in Mathematics, Boğaziçi University, 2014

M.S., in Mathematics, Boğaziçi University, 2017

Submitted to the Institute of Biomedical Engineering

in partial fulfillment of the requirements

for the degree of

Doctor

of

Philosophy

Boğaziçi University

2021

ACKNOWLEDGMENTS

Throughout my Ph.D. studies, I have received a great deal of encouragement and support.

I would first like to express my very sincere thanks to my advisor Assoc. Prof. Dr. Bora Garipcan and co-advisor Dr. Mehmet Turan for their consistent motivation, and support. I appreciate their feedback, which always bring my studies to the higher level.

I would like to thank to my thesis progress committee members, Prof. Dr. Cengizhan Öztürk and Prof. Dr. Erkan Kaplanoğlu, for their thoughtful comments and invaluable suggestions to improve the quality of my thesis work. I would like to express my gratitude to my other thesis committee members, Assoc. Prof. Dr. Esin Öztürk Işık and Dr. Ali Bayram for allocating time to review my thesis, their contributions and thoughtful comments.

ACADEMIC ETHICS AND INTEGRITY STATEMENT

I, Kutsev Bengisu Özyörük, hereby certify that I am aware of the Academic Ethics and Integrity Policy issued by the Council of Higher Education (YÖK) and I fully acknowledge all the consequences due to its violation by plagiarism or any other way.

Name :

Signature:

Date:

ABSTRACT

Deep Learning Approaches for the Localization of Capsule Endoscope

Deep learning techniques hold promise to develop dense topography reconstruction and pose estimation methods for endoscopic videos. However, currently available datasets do not support effective quantitative benchmarking. In this thesis, we introduce a comprehensive endoscopic simultaneous localization and mapping (SLAM) dataset consisting of 3D point cloud data for six porcine organs, capsule and standard endoscopy recordings, synthetically generated data as well as clinically in use conventional endoscope recording of the phantom colon with computed tomography scan ground truth. To verify the applicability of this data for use with real clinical systems, we recorded a video sequence with a state-of-the-art colonoscope from a full representation silicon colon phantom. Additionally, we propound Endo-SfMLearner, an unsupervised monocular depth and pose estimation method that combines residual networks with a spatial attention module in order to dictate the network to focus on distinguishable and highly textured tissue regions. The proposed approach makes use of a brightness-aware photometric loss to improve the robustness under fast frame-to-frame illumination changes that are commonly seen in endoscopic videos. To exemplify the use-case of the EndoSLAM dataset, the performance of Endo-SfMLearner is extensively compared with the state-of-the-art: SC-SfMLearner, Monodepth2, and SfMLearner.

Keywords: SLAM Dataset, Capsule Endoscopy, Spatial Attention Module, Monocular Depth Estimation, Visual Odometry.

ÖZET

Kapsül Endoskopi Lokalizasyonu İçin Derin Öğrenme Teknikleri

Derin öğrenme teknikleri endoskopi vidyolarında yoğun topografi yeniden canlandırma ve lokasyon tahmini methodları için ümit vaad etmektedir. Ancak, şuan anonim veri kümeleri efektif sayısal kıyaslamayı desteklememektedir. Bu tezde, altı domuz iç-organı ile eş-güdümlü konumlandırma ve haritalandırma algoritmaları geliştirmede kullanılabilecek 3D nokta bulutu datası, kapsül ve standart endoskopi kayıtları oluşturuldu. Ayrıca, Unity ortamında sentetik olarak üretilmiş ve standart klinik kullanımdaki endoskop ile fantom kolondan toplanan bilgisayarlı tomografi taramasını kesin referans olarak içeren veri eklenerek kapsamlı bir endoskopi dataset oluşturulmuştur. Buna ek olarak, Endo-SfMLearner, konumsal dikkat modulü ile derin kalıntı ağlarını kombinleyen güdümsüz monokülar derinlik ve pozisyon tahmini methodu önerilmiştir. Parlaklık farkındalıklı fotometrik yitim fonksiyonu sayesinde endoskopik vidyolarda sıkça görülen kamera kareleri arası hızlı ışık değişimlerine karşı dayanıklılık artırılmıştır. EndoSLAM veri kümesi kullanımı, Endo-SfMLearner algoritmasının en yaygın kullanılan methodlarla; SC-SfMLearner, Monodepth2 ve SfMLearner ile geniş kıyaslaması ile örneklenmiştir.

Anahtar Kelimeler: SLAM Veri Kümesi, Kapsül Endoskopi, Konumsal Dikkat Modül, Monokülar Derinlik Tahmini, Görsel Odometri.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

$\mathbf{X_g}$	point in the reference frame of the gripper
$\mathbf{X_c}$	point in the reference frame of the camera
$\mathbf{t_g^c}$	translation vector from reference frame
$\mathcal{L}_{bp}^{M}$	brightness-aware photometric loss
$\mathcal{L}_s$	smoothness loss
$\mathcal{L}_{GC}$	geometry consistency loss
u	distance to subject and camera
f	focal length
c	circle of confusion
N	the ratio of the systems' focal length to the diameter of entrance pupil
α	Gaussian Blur filter size
β	Gaussian Blur standard deviation
γ	the number of Gaussian Blur filtering times
Δ	fixed trajectory length

LIST OF ABBREVIATIONS

GI	Gastrointestinal
SLAM	Simultaneous Localization and Mapping
DVI	Direct Visual Inspection
EGD	Esophagogastroduodenoscopy
CE	Capsule Endoscopy
VV	Vestre Viken
ESAB	Endosfmlearner Spatial Attention Block
MDSS	The Medical Decision Support Systems
WCE	Wireless Capsule Endoscopy
NBI	Narrow Band Imaging
CNN	Convolution Neural Network
EKF-SLAM	Extended Kalman Filter Simultaneous Localization and Mapping
PTAM	Parallel Tracking and Mapping
CT	Computed Tomography
vSLAM	visual Simultaneous Localization and Mapping
RCNN	Recurrent Convolutional Neural Network
EndoAbS	The Endoscopic Abdominal Stereo Images
DoF	Depth of Field
dof	Degree of freedom
BN	Batch Normalization
ATE	Absolute Trajectory Error
RMSE	Root Mean Square Error
SIFT	Scale Invariant Feature Transform
RPE	Relative Pose Error
ICP	Iterative Closest Point

1. INTRODUCTION

1.1 Motivation

Gastrointestinal (GI) cancers affect over 28 million patients annually, representing about 26% of the global cancer incidence and 35% of all cancer-related deaths [2]. Besides, GI cancer is the second deadliest cancer type with reported 3.4 million GI related deaths globally in 2018 [3]. Direct visual inspection (DVI) of these cancers is the simplest and most effective technique for screening. Esophagogastroduodenoscopy (EGD) and colonoscopy are used to visualize gastrointestinal diseases in colon and rectum while capsule endoscopy (CE) is preferred for small bowel exploration [4].

An endoscopic gastro-intestinal procedure analysis held by iData Research reveals that over 19 million colonoscopies are performed annually, as reported in 2017; a tremendous contribution to the 75 million endoscopies performed each year in the United States [5]. Specifically, the malignant tumors developed in the small intestine like Adenocarcinoma, Intestinal Lymphoma, Leiomyosarcoma, and metastatic malignancy from lung or breast are severe diseases, mostly resulting in death. Among these, the small bowel involving polyposis syndromes include Familial Adenomatous Polyposis, generalized Juvenile polyposis, Peutz-Jeghers, and Cronkhite-Canada syndromes are the most mortal types. The diagnosis of these polyps and small-bowel tumors are challenging due to the rarity of lesions, lack of common symptoms across patients, and variety of the symptoms [6]. In these cases, differential diagnosis from blood tests and symptoms alone are not sufficient, and visual examination through capsule endoscopy can provide valuable information. After visual confirmation of any feature of diagnostic importance, "Where is it?" arises as a natural question. In the following subsection, we overview the related work from the literature which are all motivated by this critical question.

1.2 Objectives and Outline of the Thesis

In this work, we introduce the EndoSLAM dataset, a dedicated dataset designed for the development of 6-dof pose estimation and dense 3D map reconstruction methods. The dataset is recorded using multiple endoscope cameras and ex-vivo porcine GI organs belonging to different animals and is designed to meet the following major requirements for scientific research and development of endoscopic SLAM methods:

- Time-synchronized, ground-truth 6-dof pose data
- High precision, ground-truth 3D reconstructions
- Multiple organs from multiple individuals
- Images from cameras with varying intrinsic properties
- Image sequences with differing native frame rates
- Images acquired from different camera view angle such as perpendicular, vertical and tubular
- Images under a variety of lighting conditions
- Distinguishable features of diagnostic significance (e.g. presence/absence of polyps).

In addition to the experimentally collected data, synthetically generated data from a 3D simulation environment is included to facilitate the study of the simulation to real-world problems such as domain adaptation and transfer learning. One of the biggest disadvantages of deep learning techniques is the fact that large networks need massive amounts of domain-specific data for training. Research in recent years has shown that large amounts of synthetic data can improve the performance of learning-based vision algorithms and can ameliorate the difficulty and expense of obtaining real data in a variety of contexts. However, due to the large gap between simulation data and real data, this path needs domain adaptation algorithms to be employed. With the

synthetically generated data from Unity 3D environment, we aim to provide a test-bed to overcome the gap between simulation and real endoscopic data domain.

In addition to the EndoSLAM dataset, we propose an unsupervised depth and pose estimation approach for endoscopic videos based on spatial attention and brightness-aware hybrid loss. The main idea and details of the proposed architecture are depicted in Fig. 4.1. Our main contributions are as follows:

- Spatial Attention-based Visual Odometry and Depth-Estimation: We propose spatial attention based ResNet architecture for pose estimation optimized for endoscopic images.
- Hybrid Loss: We propose a hybrid-loss function which is specifically designed to cope with the depth of field related defocus issues and fast frame-to-frame illumination changes in endoscopic images. It collaboratively combines the power of brightness-aware photometric loss, geometry consistency loss, and smoothness loss.

2. BACKGROUND

2.1 SLAM Dataset

In the literature, various open source dataset are released to support the development of localization, disease detection, 3D-map reconstruction and depth estimation algorithms. We are over-viewing some of them as below.

- The KID Dataset is organized by The Medical Decision Support Systems (MDSS) research group of the University of Thessaly. The dataset is divided into two annotated sections. The first section has a total of 77 wireless capsule endoscopy (WCE) images acquired using MiroCam® (IntroMedic Co, Seoul, Korea) capsules and has some types of abnormalities such as angioectasias, apthae, chylous cysts and polypoid lesions. The second part consists of 2,371 MiroCam® WCE. This dataset not only includes small bowel lesions such as polypoid, vascular and inflammatory lesions but also images from healthy esophagus, stomach, small bowel and colon Given Imaging Atlas Dataset consists of 20 second video clips recorded using PillCam capsules with a resolution of 576x576 pixels. In this database, 117 WCE video clips have been acquired from the small bowel, 5 from esophagus and 13 from the colon [7].

- The Kvasir dataset was collected via standard endoscopic equipments at Vestre Viken (VV) Health Trust in Norway. The initial dataset consists of 4,000 images with eight classes namely Z-line, pylorus, cecum, esophagitis, polyps, ulcerative colitis, dyed and lifted polyps and dyed resection margins of images, each represented with 500 images. All images are annotated and verified by experienced endoscopists [8]. Later, the dataset extended to 8,000 images with the same eight classes [9]. The Kvasir-SEG Dataset is an extension of the Kvasir dataset which is used for polyp segmentation. It comprises 1000 polyp images and their corresponding ground truth from the second version of the Kvasir dataset [10].

Table 2.1
Dataset survey. An overview of existing datasets for disease classification, polyp recognition, segmentation, pose tracking, and depth estimation. The size of each dataset in terms of the number of images and corresponding organs are also listed. The datasets, collected via capsule endoscopy, standard endoscopy, and laparoscopy are denoted by $^{\diamond}$, † and *, respectively.

Dataset Name	Findings	Organs	Tasks	Size
Kvasir-SEG†	Polyps	Colon	Segmentation	1,000
Kvasir† [8]	Z-line, pylorus, cecum, esophagitis, polyps, ulcerative colitis, dyed	Colon	Disease detection	6,000
	Lifted polyps and dyed resection margins	Colon	Segmentation	2,000
Hamlyn Centre Datasets†*	Polyp	Colon	Segmentation	7,894
	-	Kidney	Disparity	40,000
	Polyp	Colon	Polyp recognition Localisation	2,000
	-	Liver, ureter, kidney, abdomen	Tissue deformation Tracking	-
KID Dataset$^{\diamond}$ [11]	Angioectasias, apthae, chylous cysts and polypoid, vascular and inflammatory lesions	Small Bowel and colon	Classification	2,448
NBI-InFrames† [12]	Angioectasias, apthae, chylous cysts and polypoid	Larynx	Classification	720
EndoAbs† [13]	-	Liver, kidney, spleen	Classification	120
ASU-MAYO Clinic†	Polyp	Colon	Segmentation	22,701
ROBUST-MIS Challenge*	Rectal cancer	Abdomen	Segmentation	10,040

- The Hyper-Kvasir dataset is the largest online available dataset related to the gastrointestinal tract, containing 110,079 images (10,662 labeled and 99,417 unlabeled images) and 373 videos, making a total of 1.17 million frames. The entire dataset was collected in gastro- and colonoscopy examinations in Norway and 10,662 images are labeled for 23 classes by practitioners. [9].

- The NBI-InfFrames dataset includes Narrow-band imaging(NBI) endoscopy which

is commonly used as a diagnostic procedure to examine the back of throat, glottis, vocal cords and the larynx. To generate this in vivo dataset, 18 different patients affected by laryngeal spinocellular carcinoma (diagnosed after histopathological examination) were involved. It consists of 180 informative (I), 180 blurred (B), 180 with saliva or specular reflections (S) and 180 underexposed (U) frames with a total number of 720 video frames [12].

- The Endoscopic Abdominal Stereo Images(EndoAbS) Dataset consists of 120 sub-datasets of endoscopic stereo images of abdominal organs (e.g., liver, kidney, spleen) with corresponding ground truth acquired via laser scanner. In order to create variations in the dataset, frames have been recorded under 3 different lighting conditions, presence of smoke and 2 different distances from endoscope to phantom ($\sim$5 cm and $\sim$10 cm). The main purpose of generating this dataset was to validate 3D reconstruction algorithms for the computer assisted surgery community [13].

- CVC-ColonDB is a database of annotated video sequences consisting of 15 short colonoscopy sequences, where one polyp has been shown in each sequence. There are 1,200 different images containing original images, polyp masks, non-informative image masks and contour of polyp masks. It can be used for assessment of polyp detection [14].

- MICCAI 2015 Endoscopic Vision Challenge [15] provides three sub-databases which are CVC-ClinicDB, ETIS-Larib and ASU-Mayo Clinic polyp database and which can be used for polyp detection and localization. CVC-ClinicDB is a cooperative work of the Hospital Clinic and the Computer Vision Center, Barcelona, Spain. It contains 612 images from 31 different sequences. Each image has its annotated ground truth associated, covering the polyp [16]. ETIS-Larib is a database consisting of 300 frames with polyps extracted from colonoscopy videos. Frames and their ground truths are provided by ETIS laboratory, ENSEA, University of Cergy-Pontoise, France [17]. The ASU-Mayo Clinic polyp database was acquired as a cooperative work of Arizona State University and Mayo Clinic, USA. It consists of 20 short colonoscopy videos (22,701 frames) with different resolution ranges and different area coverage values for training purposes. Each

frame in its training dataset comes with a ground truth image or a binary mask that indicates the polyp region. In addition, it contains 18 videos without annotation for testing purposes [18].

- The Hamlyn Centre Laparoscopic/Endoscopic Video Dataset consists of 37 subsets. The Gastrointestinal Endoscopic Dataset includes 10 videos and consists of 7,894 images with a size of 2.5 GB which were collected during standard gastrointestinal examinations. The dataset includes images for polyp detection, localization and optical biopsy retargeting. Apart from endoscopy dataset for depth estimation, one of the laparoscopy datasets contains ∼40,000 pairs of rectified stereo images collected in partial nephrectomy using Da Vinci surgery robot. Its primary use has been training and testing deep learning networks for disparity (inverse depth) estimation [19], [20].

- ROBUST-MIS Challenge provides a dataset which was created in the Heidelberg University Hospital, Germany during rectal resection and proctocolectomy surgeries. Videos from 30 minimal invasive surgical procedures with three different types of surgery and extracted 10,040 standard endoscopic image frames from these 30 procedures performed a basis for this challenge. These images were acquired using a laparoscopic camera (Karl Storz Image 1) with a 30° optic and a resolution of 1920x1080 pixels. The images are, then, downscaled to 960x540 pixels and annotated with numbers showing the absence or presence of medical instruments [21].

2.2 Visual Odometry and Depth Estimation

The direction of arrival estimation based localization techniques such as radio frequency based signal triangulation [22], received signal strength [23], electromagnetic tracking [24], x-ray [25] and positron emission markers [26] have been widely investigated in robotics. In capsule endoscopes, visual information has been provided which has driven attention to the development of vision-based odometry and simultaneous localization and mapping (SLAM) systems, either to remove the need for added hard-

ware for pose sensing or to provide additional information for 3D tracking. While current capsules are propelled by the peristaltic motion of the GI tract, active capsule endoscopes hold promise to provide drug delivery and biopsy [27]. Vision-based SLAM is of utmost importance to enable these functions and other forms of complementary situational awareness in decision support and augmented reality systems [28]. With the rise of deep learning techniques [29], public datasets enabling a broader research community to work on the localization and mapping problems became crucial [8, 15, 9] in medical image analysis. Several datasets are available to support research and development of a variety of advanced diagnostic features across a wide range of tasks, including segmentation, disease classification, tissue deformation and motion detection, and depth estimation. Some of them are available in the context of endoscopy datasets which are overviewed in Table 2.1 Depth estimation from a camera scene and visual odometry are very challenging and active problems in computer vision. Various traditional multi-view stereo [30] methods such as structure from motion [31, 32] and SLAM [33] can be used to reconstruct a 3D map based on the feature correspondence. However, their performances are still far from being perfect especially for endoscopic images suffering from lack of distinguishable features. Despite the recent advances in image processing, colonoscopy remains as a complicated procedure for depth estimation because of the monocular camera with an insufficient light source, limited working area and frequently changing environment due to the contractions of muscles. In that regard, deep-learning based methods have been applied for monocular depth estimation [34, 35, 36]. CNN-based depth estimation methods have shown promising performance on a single image depth inference despite the scale inconsistency [37]. Nevertheless, using CNN in a fully supervised manner is challenging for endoscopy since dense depth map ground truth that corresponds directly to the real endoscopic image is hard to obtain. Even if the labeled dataset is provided, patient-specific texture, shape, and color make difficult to get generalizable results without a large amount of ground truth. These issues are mostly overcome by either synthetically generated data or the simultaneous depth and pose estimation methods where the output of pose network supervises the depth network instead of human expert annotations [38, 39]. Mahmood et al. propose an unsupervised reverse domain adaptation framework to avoid these annotation requirements which is accomplished by adversarial training removing patient specific

details from real endoscopic images while protecting diagnostic details [40]. In [41], the monocular depth estimation is formulated as conditional random fields learning problem and CNN-CRF framework that consists of unary and pairwise parts are introduced as domain adaptable approach. Several self-supervised methods related to the single-frame depth estimation have been propounded in the generic field of computer vision [42, 43, 44]. However, they are not generally applicable to endoscopy because of inter-frame photometric stability assumption of these works which is broken by the frequently appearing inconsistent illumination profile in endoscopic videos. The jointly moving camera and light source cause the appearance of the same anatomy to differ substantially with varying camera poses, especially for tissue regions close to the camera surface. This might give rise to the network getting stuck in a local minimum during training, specifically for textureless regions where extracting reliable information from photometric appearance is extremely difficult [45]. There are also studies solely focusing on monocular localization problems utilized by CNN [46, 47]. Unlike traditional artificial neural networks, Turan et al. use RCNN which is able to process arbitrarily long sequences by its directed cycles between the hidden units and infer the correlative information across frames [29]. However, estimating a global scale from monocular images is inherently ambiguous [35]. Despite all efforts, visual odometry is insufficient in real-time localization and vSLAM methods come on the scene as a solution which can be tested only via a comprehensive vSLAM dataset with accurate ground truths. In the work of Mountney et al., a vSLAM method based on Extended Kalman Filter SLAM (EKF-SLAM) is used for localization and soft tissue mapping where sequential frames are acquired by moving stereo endoscopes [48]. In robotic surgical systems such as da Vinci™, real-time 3D reconstruction methods have been applied and validated on phantom models [49, 50]. Lin et al. adopt and extend the Parallel Tracking and Mapping (PTAM) method to detect deformations on a non-rigid phantom to create 3D reconstruction of an intestine model and to track endoscope position and orientation [50]. Some other works are focused on more commonly used monocular endoscopes. Mirota et al. generate a 3D reconstruction from endoscopic video during sinus surgeries by using feature detection and registered data from computed tomography (CT) scan tracking endoscope location [51]. In [33], another monocular vSLAM method is used to provide real-time 3D map of the abdominal cavity for hernia repair interven-

tions. Apart from standard endoscopes, vSLAM techniques have also been used in capsule endoscopy [52, 53]. A robust and reliable SLAM module is indispensable for next-generation capsule robots equipped with the functionalities including biopsy, drug delivery, and automated polyp detection [54], but several technical challenges such as low frame rate and low resolution due to space limitations make this need tough to meet. Specular reflections from extracellular fluids and rapidly changing environment due to peristaltic motions are further examples of inherent challenges. Those problems have motivated the exploration of deep learning based approaches that eschew complex physical models which ends up with the necessity of a huge amount of dataset.

3. SIMULTANEOUS LOCALIZATION AND MAPPING DATASET (ENDOSLAM)

3.1 Dataset Shooting

In this section, we introduce the experimental setup, data collection procedure and the detailed structure of EndoSLAM dataset.

3.1.1 Experimental Setup

The experimental setup was specifically designed to support the collection of endoscopic videos, accurate 6-dof ground truth pose, organ shape, and topography data, see Fig. 3.1 for the illustration of all equipment. The essential components are five endoscope video cameras, a robotic arm to track the trajectory to quantify the pose values, high resolution CT scanner and high precision 3D scanners for ground truth organ shape measurement and full Chamberlian Colon Phantom. As per camera devices, MiroCam® and Pillcam® COLON2 capsule endoscope cameras, three other cameras representative of conventional endoscope cameras were employed. Their specifications are as follows:

- MiroCam® Regular MC1000-W endoscopic video capsule: 320×320 image resolution, 3 fps frame rate, 170° field of view, 7 - 20 mm depth of field, 6 white LED's, [55], Fig. 3.1 p.
- Pillcam® COLON2 double endoscope camera capsule: 256×256 each camera, 4 fps to 35 fps variable frame frate, 344° field of view (172° each camera), 4 LEDs (each camera), [55], Fig. 3.1 o.
- High Resolution Endoscope Camera (YPC-HD720P): 1280×720 image resolution, 20 fps frame rate, 120° field of view, 4-6 cm depth of field, 6 adjustable white

LEDs, Fig. 3.1 k.

- Low Resolution Endoscope 3 in 1 Camera: 640×480 image resolution, 20 fps frame rate, 130° field of view, 3-8 cm depth of field, 6 adjustable LEDs, Fig. 3.1 l.

- Olympus CFHQ-190L colonoscope, 1350×1080 image resolution, 30 fps frame rate, 170° field of view, 5-100 mm depth of field, CV-190 video processor, and CV-190L light source, Fig. 3.1 n.

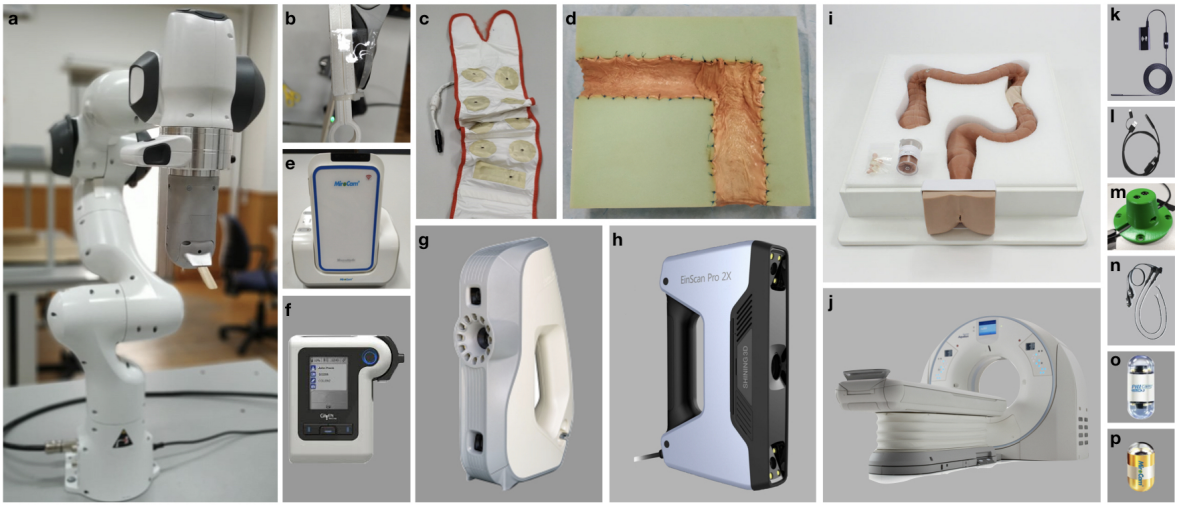


Figure 3.1 Equipment. The overall equipment for dataset generation. **a** Franka Emika Panda: motion control device for cameras. **b** Capsule Holder: two-piece holder as a kit between the WCE cameras and the robotic arm. **c** MiroCam® Data Belt **d** Real Porcine Colon: sewn onto an 'L' shaped semi-cylindrical scaffold in high-density foam. **e** MiroCam® MR1100 receiver: Digital video grabber for the conversion of analog data into digital and output to the computer. **f** PillCam® recorder **g** Artec Eva: 3D scanner used to generate ground truth - ply file. **h** EinScan Pro 2X: 3D scanner used to generate ground truth - .ply, .obj, .stl and .ASC file. **i** Full Chamberlain Colon Phantom **j** Canon Aquilion Precision CT Scanner **k** Wireless Endoscope Camera (YPC-HD720P): high resolution - 1280Ã720 and HD640Ã480. **l** Endoscope 3 in 1 Camera: low resolution - 640Ã480. **m** Camera Holder: specially designed one-piece holder for the stabilization of the high and low resolution endoscope to the robotic arm. **n** Olympus CFHQ-190L colonoscope **o** PillCam™ COLON2: WCE double tip camera. **p** MiroCam® Regular MC1000-W: WCE camera.

Ground truth geometries of the organs were acquired via Canon Aquilion Precision CT scanner as well as two commercially-available 3D scanners, the Artec 3D Eva and Shining 3D Einscan Pro 2x. 3D models of organs were reconstructed as in Fig. 3.2 and the depth distribution histograms for corresponding organs are given in Fig. 3.8. Relevant performance specifications of the CT and 3D scanners are as follows:

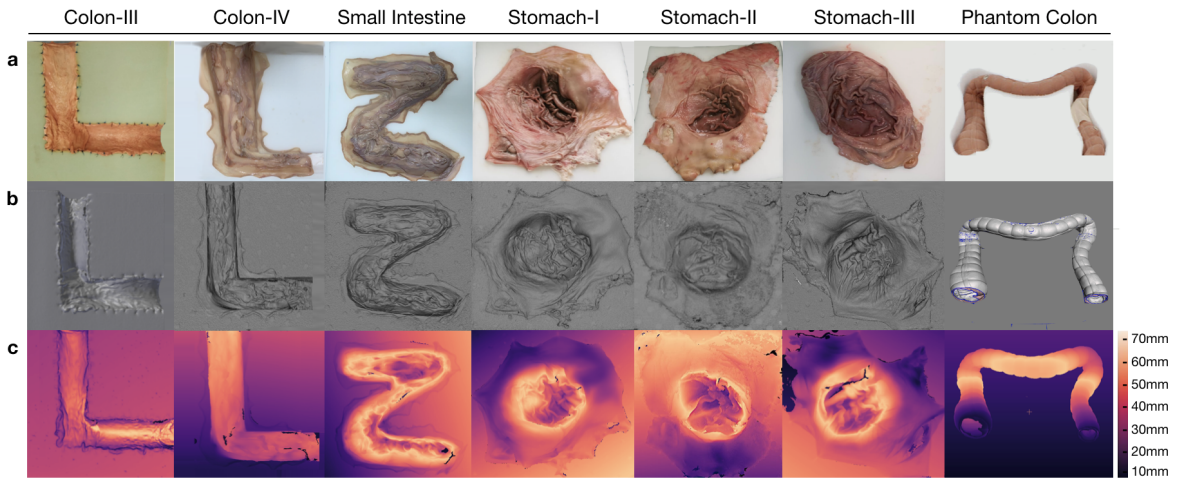


Figure 3.2 3D-Scanner images for ex-vivo organs and CT-Scanner image for colon phantom. 3D-scanner images for six organs which are fixed to scaffolds that were cut in O, Z, and L shapes as well as colon phantom. **a** RGB images of scanned organs. **b** Corresponding 3D reconstruction from .ply files for ex-vivo organs recorded via 3D Scanner and 3D reconstruction of colon phantom from .dcm files recorded via CT scanner. **c** Heatmap reconstructions for depth values by means of the Computer Vision Toolbox of Matlab. 3D point cloud data from two colons, one small intestine, and three stomachs from different individuals make dataset appropriate not only for the development of 3D reconstruction algorithms but also for transfer learning problems.

- Canon Acquilion Precision CT Scanner: 150 micron, 50 lp/cm* resolution, 1024 matrix Ultra-High Resolution, 160 detector rows, and 1792 channels of only 0.25 mm thickness, Fig. 3.1 j.
- Artec 3D Eva: ±0.5 mm 3D resolution, ±0.1 mm 3D point accuracy, ±0.03% 3D accuracy over 100 cm distance, [56], Fig. 3.1 g.
- Shining 3D EinScan Pro 2x: 0.2-2mm point distance; ±0.5 mm 3D resolution, ±0.05 mm 3D point accuracy, ±0.03% 3D accuracy over 100 cm distance, [57], Fig. 3.1 h.

3.1.2 Synthetic Data Generation

In addition to the real ex-vivo part of the EndoSLAM dataset, we have generated synthetic capsule endoscopy frames to facilitate the study of simulation-to-real transfer of learning-based algorithms. The simulation environment, VRCaps [58], provides

synthetic data which is visually as well as morphologically realistic. The platform was built with the use of real CT images in DICOM format for topography and endoscopic images in RGB format for texture assignment. A cinematic rendering tool mimicking the effects in real capsule endoscopy records such as specular reflection, distortion, chromatic aberration, and field of view was used in order to obtain more photo-realistic images. Operating the virtual capsule inside the virtual 3D GI tract, we have recorded three sample endoscopic videos that containing 21,887 frames from colon, 12,558 frames from small intestine, and 1,548 frames from stomach with pixel size of 320x320 and having both positional and pixel-wise depth ground truth.

3.2 Data Tree Structure

EndoSLAM dataset is divided into four main parts: Cameras, 3D_Scanners, OlympusCam and UnityCam. Each subfolder of "Cameras" branches out into calibration and organs subfolders. Calibration subfolder comprises intrinsic-extrinsic camera parameters and corresponding calibration sessions whereas organs subfolder includes frames and camera pose ground truth of each trajectories. 3D_Scanners folder consists of reconstructed 3D figures (.fig), point cloud data (.ply), surface geometry of three-dimensional objects without any color or texture representations (.STL), the position of each vertex representing 3D geometry(.obj) and ASCII formatted point cloud data(.ASC). OlympusCam folder includes calibration parameters, the rgb frames from phantom colon and CT scan ground truth. Finally UnityCam folder includes synthetically generated images, pixelwise depths and corresponding 6D poses.

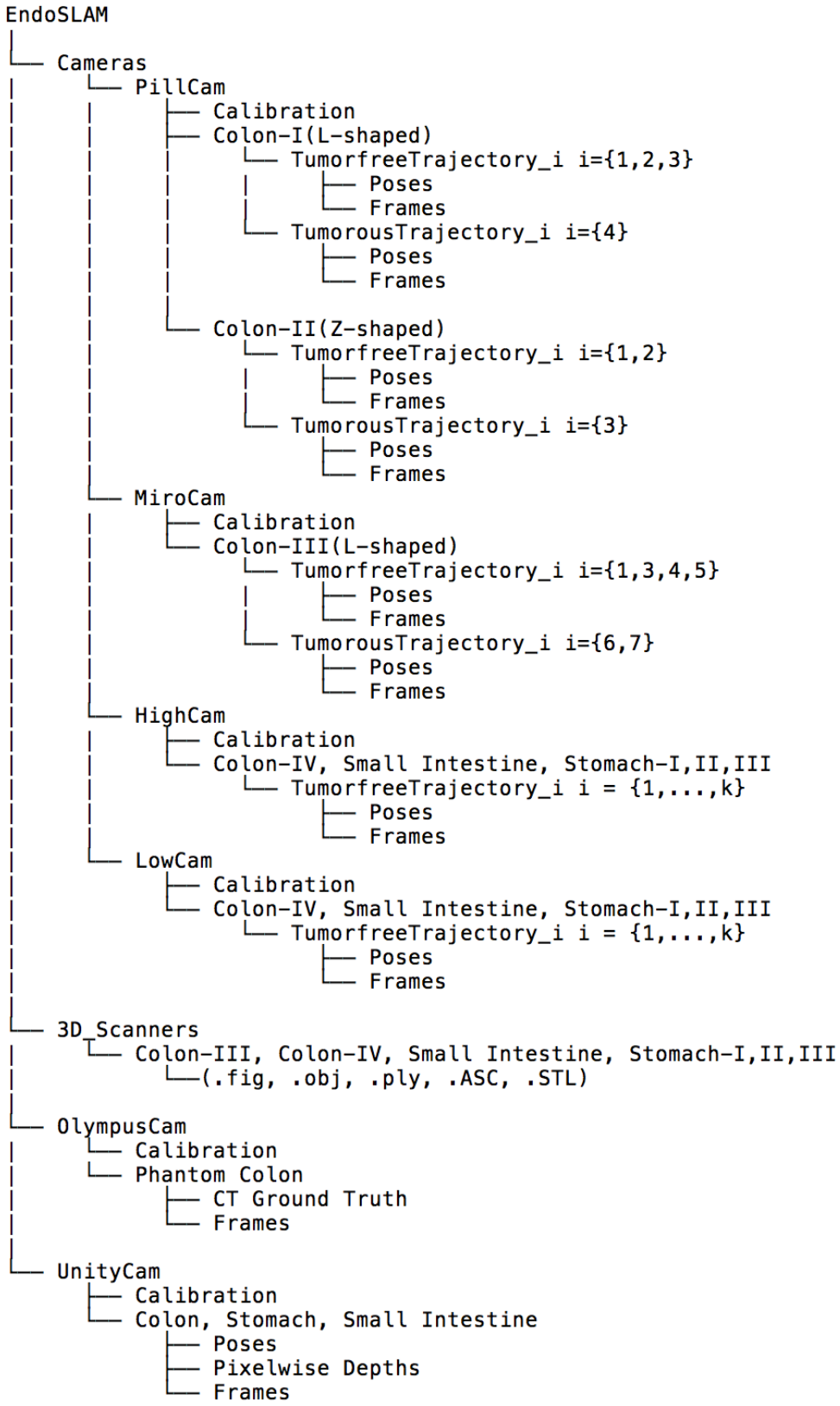


Figure 3.3 Data tree. Overall architecture of EndoSLAM dataset.

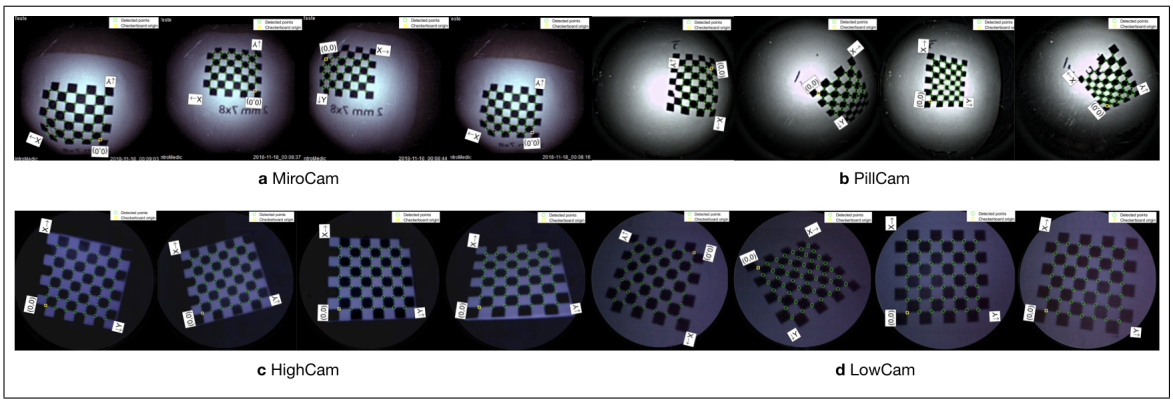
3.3 Calibration

3.3.1 Camera Calibration

The intrinsic parameter calibration was performed for both the Mirocam and Pillcam capsules, using images of a planar checkerboard with 8×7 squares of dimension 2×2 mm and also for HighCam and LowCam using 8×7 squares of dimension 12.8×12.8 mm. The calibration checkerboard was printed using a laser printer and then glued on the surface of a glass plate to ensure the planarity of the pattern.

a Mirocam pinhole **b** Pillcam Cam1 pinhole **c** Pillcam Cam2 pinhole **d** Mirocam fisheye **e** Pillcam Cam1 fisheye **f** Pillcam Cam2 fisheye

Figure 3.4 Reprojection errors associated with the camera calibrations. The reprojection errors under pinhole camera assumption for **a** Mirocam, **b** Pillcam with a front-facing (Cam1) **c** Pillcam with a backwards-facing (Cam2) camera. **d-f** Reprojection errors for the same devices under the fisheye model assumptions.



a MiroCam **b** PillCam **c** HighCam **d** LowCam

Figure 3.5 Camera intrinsic-extrinsic calibration images. Examples of planar checkerboard calibration images obtained by **a** MiroCam, **b** PillCam, **c** HighCam and **d** LowCam. The chessboards are printed with a laser printer and then glued on the surface of a planar glass to ensure the planarity of the pattern. Since the dataset is recorded in dark room, chessboard images are taken in same environmental conditions.

The practical distance and orientation range at which the calibration checkerboard can be placed is limited by the low resolution and depth of field of the cameras. For each camera, 10 calibration images were used with the pattern placed at different

Table 3.1
Intrinsic parameters for HighCam, LowCam, OlympusCam. Each camera was calibrated against a pinhole camera model with non-linear radial lens distortion by Camera Calibration Toolbox MATLAB R2020a based on the theory of Zhang [1] with the chessboard images illustrated in Fig. 3.5.

			HighCam	LowCam	OlympusCam
		H x W	480×640	480×640	1080×1350
Pinhole	Focal length	f_x	957.4119	816.8598	768.2788
		f_y	959.3861	814.8223	769.8207
	Skew	s	5.6242	0.2072	1.0464
	Optical center	c_x	282.1921	308.2864	676.9603
		c_y	170.7316	158.3971	540.0451
	Radial dist. coef.	k_1	0.2533	0.2345	-0.4933
		k_2	-0.2085	-0.7908	0.2531

poses. The average distance from the camera was approximately 10 mm for capsule cameras. Fig. 3.5 show examples of some of the calibration images.

3.3.2 Hand-Eye Calibration

For the coordinate transformation between robot pose data and capsule cameras, hand-eye calibration procedure was repeated with two different checkerboards: one with 2×2 mm squares and one with 1.5×1.5 mm squares, both patterns with 8×7 squares in total. Four images of each checkerboard were acquired from different camera poses. For the pose conversions, only the checkerboard images from Mirocam capsule was used, with the support structure being the same for both capsules (Pillcam and Mirocam). Similarly, to calculate the transformation between the gripper holding HighCam-LowCam and the camera positions, same procedure was repeated by using the checkerboard with 10.2×10.2 mm squares.

Table 3.2
Intrinsic parameters for MiroCam, and PillCam. Each camera was calibrated against a pinhole camera model with non-linear radial lens distortion by Camera Calibration Toolbox MATLAB R2020a based on the theory of Zhang [1] with the chessboard images illustrated in Fig. 3.5.

				PillCam	
			MiroCam	Cam1	Cam2
		H x W	320 × 320	256 × 256	256 × 256
Pinhole	Focal length	f_x	156.0418	74.2002	76.0535
		f_y	155.7529	74.4184	75.4967
	Skew	s	0	0	0
	Optical center	c_x	178.5604	129.9724	130.9419
		c_y	181.8043	129.1209	128.4882
	Radial dist. coef.	k_1	-0.2486	0.1994	0.1985
		k_2	0.0614	-0.1279	-0.1317

The Tsai and Lenz algorithm[59] was tested with 24 combinations of the 4 checkerboard images in Fig. 3.5. The transformation between a point $\mathbf{X_c}$ in the reference frame of the camera and a point $\mathbf{X_g}$ in the reference frame of the gripper is given by

$$\mathbf{X_g} = \mathbf{R_g^c}\,\mathbf{X_c} + \mathbf{t_g^c} \tag{3.1}$$

with the rotation matrices and translation vectors given in Table 3.3.

Table 3.3
Robot pose to camera transformation. The rotation matrices and translation vectors for MiroCAM, HighCam and LowCam to apply the transformations given in Eqn. 3.1. These values are provided as a .txt file and as a .mat file in the calibration folders of the EndoSLAM Dataset.

Camera	Rotation $\mathbf{R_g^c}$	Translation $\mathbf{t_g^c}$(mm)
MiroCam	$\begin{bmatrix} -0.9366 & -0.3242 & -0.1325 \\ 0.1738 & -0.1017 & -0.9795 \\ 0.3041 & -0.9405 & 0.1516 \end{bmatrix}$	$\begin{bmatrix} 2.9793 \\ -27.0224 \\ 72.1070 \end{bmatrix}$
HighCam	$\begin{bmatrix} 0.9463 & -0.0921 & -0.3098 \\ -0.1389 & 0.7495 & -0.6472 \\ 0.2918 & -0.6555 & 0.8965 \end{bmatrix}$	$\begin{bmatrix} -46.2017 \\ 20.9074 \\ 94.6349 \end{bmatrix}$
LowCam	$\begin{bmatrix} 0.8294 & 0.5577 & 0.0322 \\ -0.5586 & 0.8286 & 0.0379 \\ -0.0056 & -0.0495 & 0.9988 \end{bmatrix}$	$\begin{bmatrix} 6.0169 \\ 39.5114 \\ 101.6431 \end{bmatrix}$

3.4 Temporal Synchronization

After taking the positional and visual records, matching frames with correct 6D-positional data is extremely challenge.

Figure 3.6 Correction of lens distortions. Examples to correct the lens distortions via camera parameters given in Table 3.2 for the images acquired by PillCam and MiroCam. **a** Original 8×7 checkerboard image with 2×2mm squares obtained by PillCam, **b** Undistorted checkerboard image with pinhole calibration parameters, **c** Undistorted checkerboard image with fisheye parameters, **d** Newspaper image which is rich in texture details taken by frontal camera of PillCam **e** Undistorted counterpart of newspaper image with the calculated parameters under fisheye camera assumption. Similarly, **f** Original Colon-III image of MiroCam and **g** Undistorted version by the parameters of fisheye calibration model.

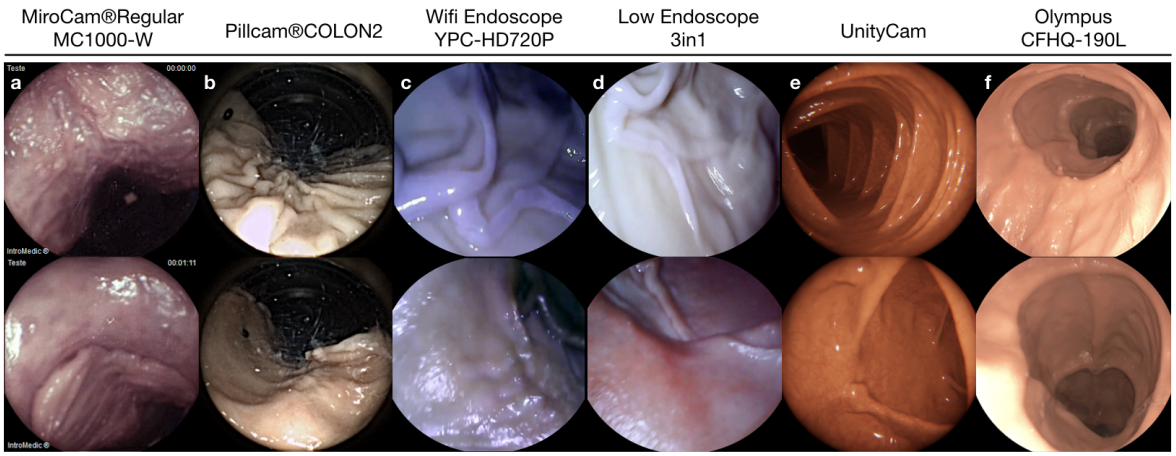


Figure 3.7 Sample frames from EndoSLAM Dataset. Images are acquired by **a** MiroCam capsule endoscope, **b** Frontal camera of a PillCam, **c** HighCam, **d** LowCam, **e** virtually generated UnityCam, and **f** OlympusCam. The ex-vivo part of the dataset offers an opportunity to test the robustness of pose estimation algorithms with images coming from various endoscope cameras. Since EndoSLAM dataset contains real and simulated frames, it is also a suitable platform to develop domain adaptation algorithms.

Table 3.4
Temporal synchronization. Correspondence, for each sequence of each organ, between the first frame of the trajectory for both HighCam and LowCam and the matching sample instant of the robot data with 1kHz recording frequency.

		Camera		Robot	
Organ	Trajectory	HighCam Start Frame	LowCam Start Frame	HighCam Sample	LowCam Sample
Colon-IV	1	741	393	35,295	15,845
	2	44	128	2,561	2,561
	3	69	82	3,975	3,975
	4	138	120	15,792	15,092
	5	99	144	1,270	3,270
SmallIntestine	1	149	95	5,162	4,512
	2	133	112	4,913	2,763
	3	186	144	6,095	7,845
	4	121	79	3,205	3,205
	5	138	105	3,807	3,307
Stom-I	1	60	135	4,443	8,093
	2	111	144	4,177	2,277
	3	71	447	6,058	19,008
	4	47	316	2,839	13,289
Stom-II	1	255	125	9,641	5,141
	2	1	2	3,358	3,358
	3	150	83	5,797	2,247
	4	78	85	2,742	4,192
Stom-III	1	195	89	6,746	2,846
	2	302	108	1,523	2,725
	3	387	105	17,261	2,861
	4	125	60	4,451	2,101

Table 3.5
Temporal synchronization. Correspondence, for each sequence, between the first frame of the trajectory and the matching sample instant (sample number) of the robot data. Note that, in the Pillcam capsule, Cam1 (front facing camera) and Cam2 (backward facing camera) trigger alternatively, one after the other, with equally spaced time intervals. The values indicated in the table correspond to Cam1.

		Camera		Robot	
Sequence		start frame	framerate	sample instant	sampl. freq.
Mirocam	1	336	3 fps	72,050	1kHz
	2	153	3 fps	961	1kHz
	3	321	3 fps	47,667	1kHz
	4	143	3 fps	33,943	1kHz
	5	254	3 fps	2,886	1kHz
	6	134	3 fps	3,044	1kHz
Pillcam	"L"	1,127	0.117 fps	15,800	1kHz
	"Z"	815	0.117 fps	11,650	1kHz

3.5 Motion Analysis

In this section, we have represented the statistical analysis of camera motions. For all trajectories of each organ, counts of robot sample instances, mean, first quantile(1st QT), median, third quantile(3rd QT), minimum, maximum speed[mm/s] values are given for HighCam in Table 3.6 and for LowCam in Table 3.7. Apart from that, the recorded trajectories for each organ divided into two groups based on the tumorous properties of tissue as tumor-containing and tumor-free as detailly given in Table 3.8.

Table 3.6
Motion analysis for HighCam. Statistics for robot poses matching with frames of **HighCam**.

		Stomach-I	Stomach-II	Stomach-III	Small Intestine	Colon-IV
Speed	frame count	4695	3302	3230	6487	3697
	mean[mm/s]	18.256	19.471	20.031	16.764	17.123
	std[mm/s]	22.497	16.809	16.697	14.210	12.660
	1st QT	5.931	7.606	7.055	5.684	7.658
	median	14.642	16.021	16.489	13.849	15.096
	3rd QT	25.32	26.64	28.68	24.342	24.324
	min[mm/s]	0.02	0.028	0.02	0	0.007
	max[mm/s]	25.32	140.898	116.984	104.08	104.759
Acceleration	mean[mm/s]	359.843	382.928	383.829	328.568	326.241
	std	450.939	337.08	336.098	284.423	257.08
	1st QT	110.408	140.982	111.71	103.129	122.807
	median	284.729	314.784	315.012	269.316	283.728
	3rd QT	501.31	528.799	556.451	477.991	469.113
	min[mm/s]	0.4	0.0	0.0	0.0	0.015
	max[mm/s]	14,680.15	2,817.962	2,339.683	2,079.994	2,095.182

Table 3.7
Motion Analysis for LowCam. Statistics for robot poses matching with frames of **LowCam**. For all trajectories of each organ, counts of robot sample instances, mean, first quantile(1st QT), median, third quantile(3rd QT), minimum, maximum speed[mm/s] values are given.

		Stomach-I	Stomach-II	Stomach-III	Small Intestine	Colon-IV
Speed	frame count	2302	2799	3900	5098	3857
	mean[mm/s]	15.599	18.928	25.97	17.918	17.144
	std[mm/s]	12.855	14.431	21.564	14.764	12.882
	1st QT	5.407	8.259	10.789	6.126	7.401
	median	13.18	15.871	21.284	15.322	15.148
	3rd QT	22.956.	26.436	35.763	26.146	24.455
	min[mm/s]	0.02	0.0	0.02	0.0	0.028
	max[mm/s]	79.042	103.254	286.68	97.315	106.271
Acceleration	mean[mm/s]	279.254	378.346	519.361	334.373	355.941
	std	253.777	288.769	431.327	295.646	259.482
	1st QT	66.573	164.972	215.786	119.299	130.256
	median	221.297	317.344	425.678	303.582	291.994
	3rd QT	428.253	528.695	715.263	520.839	482.864
	min[mm/s]	0.4	0.0	0.0	0.0	0.015
	max[mm/s]	1,580.846	2.065,071	5,733.593	1,946.305	2,125.42

Table 3.8
The classification of trajectories Approximately 10% of all trajectories is tumorous which might be practical for segmentation and disease classification tasks.

Organs	Tumor-free Trajectory #	Tumor-containing Trajectory #
Colon-I	I,II,III,	IV
Colon-II	I,III,IV,V	VI,VII
Colon-III	I,II	III
Colon-IV	I,II,III,IV,V	-
Stomach-I	I,II,III,IV	-
Stomach-II	I,II,III,IV	-
Stomach-III	I,II,III,IV	-
Small Intestine	I,II,III,IV,V	-

Table 3.9
3D-Point cloud data. The point cloud counts in 3D_Scanner folder containing six polygon (.ply) files, for which Colon-III is scanned by Artec 3D Eva with precision 0.1 mm. Colon-IV, Small Intestine and Stomach-I,-II,-III are scanned by Shining 3D EinScan Pro 2x with the precision 0.05 mm.

Organ	3D Point Count	Scanner	Precision
Colon-IV	2,106,046	3D EinScan Pro 2x	0.05 mm
Small Intestine	2,193,364	3D EinScan Pro 2x	0.05 mm
Stomach-I	2,597,906	3D EinScan Pro 2x	0.05 mm
Stomach-II	5,729,625	3D EinScan Pro 2x	0.05 mm
Stomach-III	2,234,849	3D EinScan Pro 2x	0.05 mm
Colon-III	151,846	Artec 3D Eva	0.10 mm

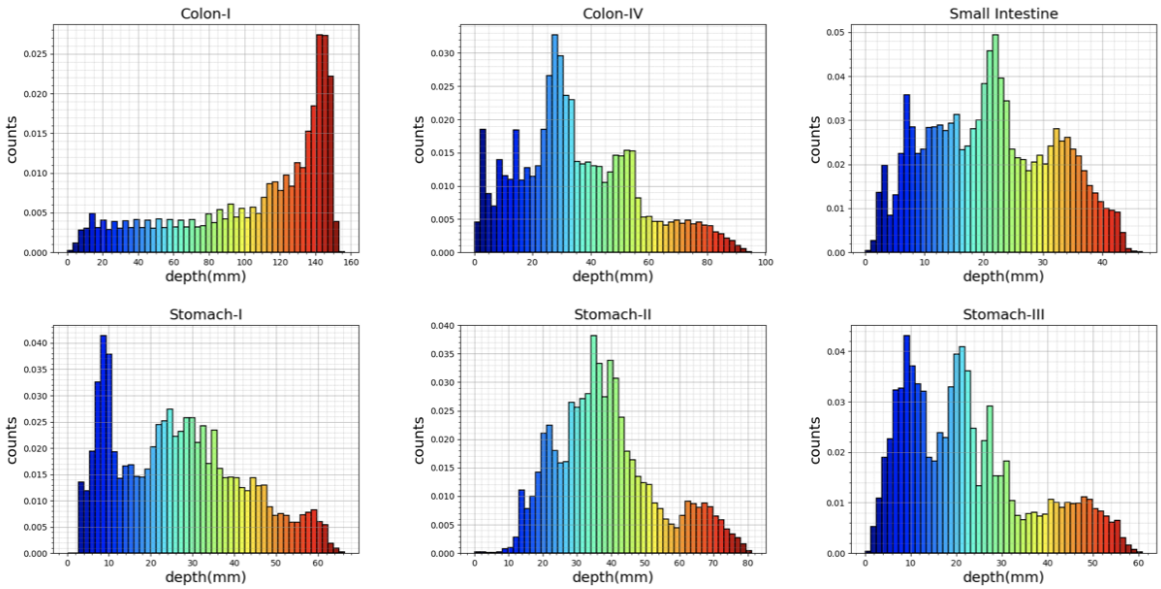


Figure 3.8 Depth evaluation of point cloud data. The frequency distribution of depth values in mm for **a** Colon-I scanned by Artec Eva: 3D scanner, **b** Colon-IV, **c** Small Intestine, **d**,**e**,**f** Stomach-I,II,III all scanned by EinScan Pro 2X.

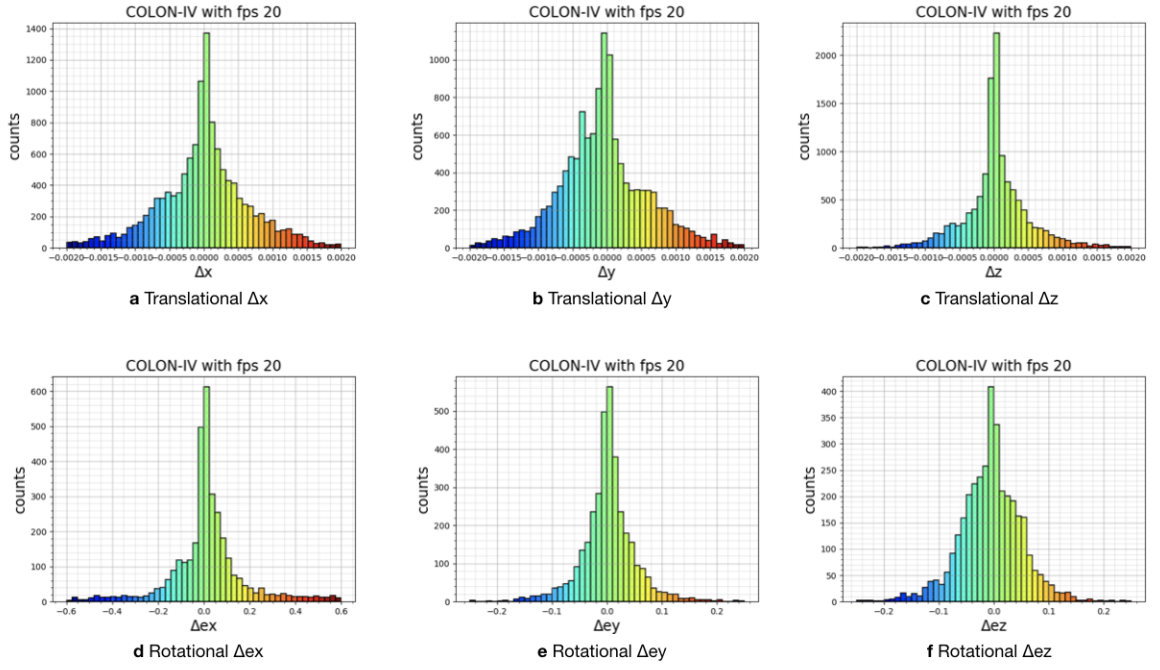


Figure 3.9 Motion analysis histograms The frequency distribution of positional differences between two consecutive frames along the **a** x, **b** y, **c** z axis and the rotational differences in **d** x, **e** y, **f** z axis in terms of Euler angles are given.

3.6 Data Augmentation

Since camera resolutions and lens properties considerably differ between capsule endoscopy designs, we added modification functions like resizing, Gaussian blur, fish-eye distortion, depth of field and vignetting effects on images to enrich the dataset and application area.

3.6.0.1 Resize. Resizing is applied with opencv-python 4.2.0.32.

3.6.0.2 Gaussian Blur. Gaussian blurred effect is implemented by convolution operation to the image f, which is defined as:

$$F(x) = \int_{-\infty}^{+\infty} f(y) e^{-\frac{(x-y)^2}{4}} \tag{3.2}$$

In principal, the convolution process assigns each pixel to a new value obtained by taking the weighted average of its neighbouring pixels where the original pixel takes the highest Gaussian value. For the various kernel sizes(α), standard deviation(β) and filtering numbers(γ), effects of Gaussian Blur function of opencv-python 4.2.0.32 library.

3.6.0.3 Fish Eye Distortion. Fish-eye lenses are ultra wide-angle lenses producing wider panoramic vision. Although fish-eye lenses help to cover larger area during limited endoscopic examination time, they may pose challenges for accurate diagnosis due to radial distortion. On the other hand, it is hard to describe radial distortion as a major concern from the perspective of standard endoscopy. To mimic the capsule endoscopy visualization challenges on the standard endoscopy images, we have implemented radial distortion effects with Pygame 1.9.6 library in Python.

3.6.0.4 Vignetting. During endoscopic procedure, gradually darkening images towards edges can be observed due to light reaching the different locations on the camera sensor at different angles. This is one of the crucial problems for capsule endoscopy with wide-angle lens which is inevitable for wide field of view [60].

3.6.0.5 Depth of Field. In the conventional and capsule endoscopy, depth of field (DoF) which is the limited and fixed distance between the nearest and farthest objects that appears clear in an image is one of the limitations of systems stems from lens properties [61]. The approximation to the DoF can be given by:

$$DoF \approx \frac{2u^2Nc}{f^2} \tag{3.3}$$

for a given distance to subject (u), focal length (f), circle of confusion (c) and the ratio of the system's focal length to the diameter of the entrance pupil (N). The effects are simulated with the shift-variant defocus bluring in MATLAB2020 [62].

3.6.0.6 Frame per second selection. The power restriction stemming from the dimensions of capsule endoscopes allows images to be transmitted only with low frame rates (∼2,3fps). This may prevent to find matched points between two consecutive frames. On the contrary, high frame rates may pose a problem to perceive the camera motion by considering consecutive frames. To observe the effects of different fps values on training and testing performance of both 3D reconstruction and visual odometry algorithms, frame selector functions can be used.

The effects of those functions on algorithm performances are shown in 3.10 and these modification functions are uploaded as open source code to our official github page.

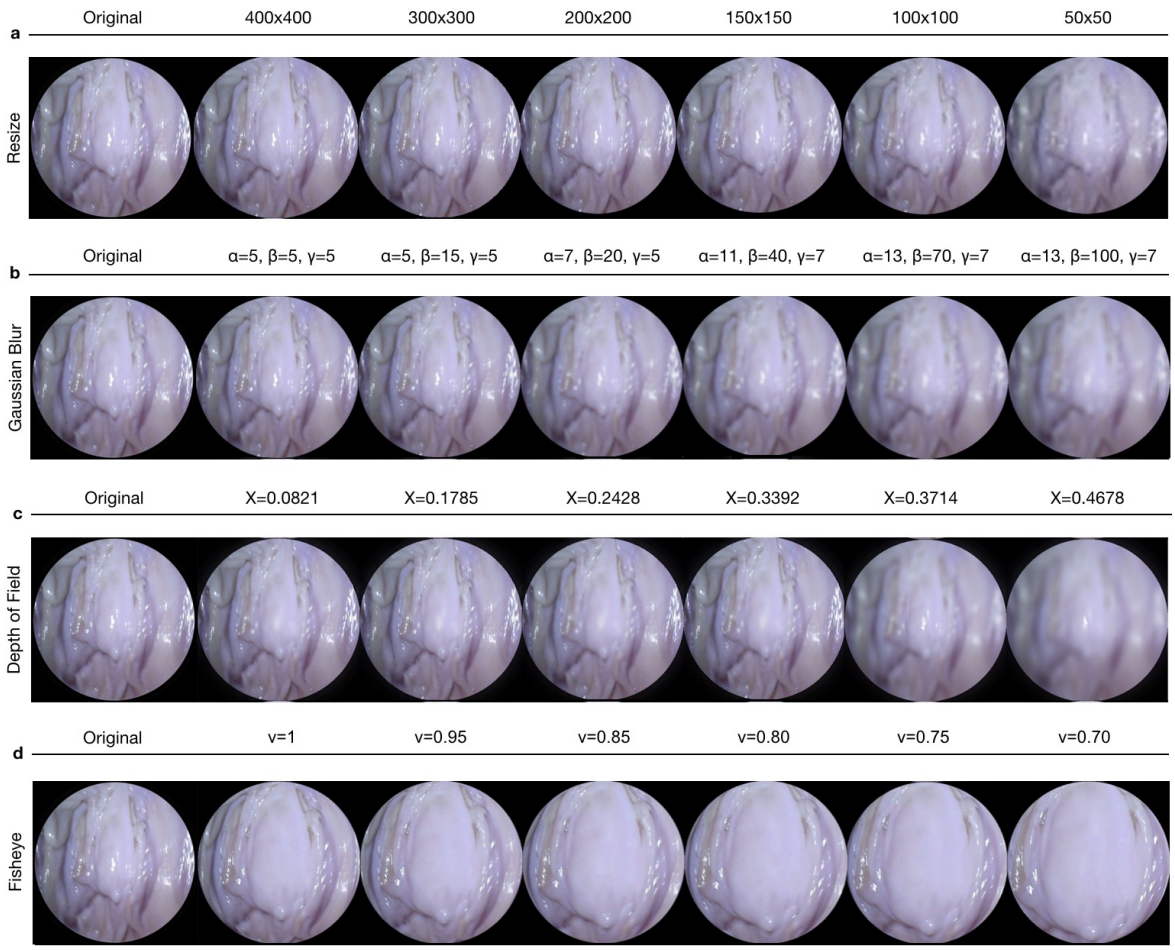


Figure 3.10 Image modifications. a Resize The size of the images, width×height, from left to right is given as 400×400, 300×300, 200×200, 150×150, 100×100 and 50×50, **b Gaussian Blur** with convolution filter size(α) are 5×5,5×5,7×7,11×11,13×13 and 13×13 and standard deviation of Gaussian distribution(β) 5,15,20,40,70,100 and the number o filtering times(γ) 5,5,5,7,7,7. **c Depth of Field** effects for the focus positions 0.0821, 0.1785, 0.2428, 0.3392, 0.3714, 0.4678, **d Fish Eye** distortion for discarding ratios ν for 1, 0.95, 0.85, 0.8, 0.75, 0.7.

4. MONOCULAR VISUAL ODOMETRY AND DEPTH ESTIMATION APPROACH FOR ENDOSCOPY VIDEOS

Recent works have proven that CNN-based depth and ego-motion estimators can achieve high performance using unlabelled monocular videos. However; static scene assumption, scale ambiguity between consecutive frames, brightness variety which basically stems from shallow depth-of-field and the organ tissues exhibiting non-lambertian surface property which are non-diffusely reflecting light particles make it difficult to provide both locally and globally consistent trajectory estimations. We are proposing the Endo-SfMLearner framework which specifically addresses these gaps.

Endo-SfMLearner jointly trains a camera pose and depth estimation networks from an unlabeled endoscopic dataset. Our method proposes two solutions to the light source rooted problems in depth and pose estimation. First proposed solution is to equate brightness conditions throughout the training and validation sets with brightness transformation function and the other is to weight the photometric loss with the brightness coefficient to punish the depth estimation with higher cost under different enlightenment conditions. Apart from these, we are using geometry consistency loss for scale-inconsistency between consecutive frames caused by alternating distances between the camera and organ tissue. In principal, we translate the estimated disparity map in one frame to 3D space, then project back into the consecutive frame via the predicted ego-motion, and decrease the inconsistency of the estimated and the projected disparity maps. This implicitly compels the depth network to produce geometrically consistent (i.e. scale-consistent) outcomes over consecutive frames. The frame-to-frame consistency will finally propagate through the whole video sequence thanks to the iterative sampling and training. Since the scale of ego-motions is strictly related to the scale of depths, the ego-motion network can estimate scale-consistent relative camera poses over consecutive pairs. The detailed network architecture for both depth and pose networks will be introduced in the following subsections.

4.1 Endo-SfMLearner Depth Network (DispNet)

Our network design is inspired and modified from previously proposed SC-SfMLearner baseline approaches [63]. The depth network which consists of encoder and decoder parts takes single image I_i as input and gives output the corresponding disparity map D_i. For the sake of brevity, hereinafter we refer to the batch normalization layer as BN, Rectified Linear Unit activation function as ReLU, exponential linear unit as ELU. Let RBk denote basic ResNet Block with k filters and `Ck` is 3x3 convolution layer with k filters. $\texttt{C}_e\texttt{k}$, $\texttt{C}_s\texttt{k}$, and $\texttt{C}_r\texttt{k}$ stand for `Ck` followed by ELU, sigmoid, and ReLU, respectively.

- **DispNet Encoder** DispNet encoder initializes with C64 with 7 kernel size, 2 stride and 3 padding followed by BN, ReLU activation function with a slope of 0.01 and max pooling operation with kernel size 3 and stride 2. Then, four ResNet basic blocks: RB64, RB128, RB256, and RB512 finalize the encoder structure. Each ResNet basic block consists of Ck(3x3), BN, ReLU, Ck, BN, and ReLU with skip connection.

- **DispNet Decoder** DispNet decoder consists of five layers each consists of two convolution operations as follows:

  $\texttt{C}_e\texttt{256}$(x2) -$\texttt{C}_e\texttt{128}$(x2)- $\texttt{C}_e\texttt{64}$(x2) - $\texttt{C}_e\texttt{32}$(x2) - $\texttt{C}_e\texttt{16}$(x2)-$\texttt{C}_s\texttt{16}$

To establish the information flow in between encoder and decoder, we are building skip connections from i^{th} to $n - i^{th}$ layer where n indicating the total number of layers and i$\in\{0,1,2,3\}$, the reader is referred to Fig. 4.1 c to overview.

4.2 Endo-SfMLearner Pose Network (Attention PoseNet)

The pose network takes the consecutive image tuples (I_i, I_{i+1}) as input by superposing and outputs the relative 6-dof pose, $P_{i,i+1}$.

- **Attention PoseNet Encoder** We have integrated attention module to the encoder of PoseNet between ReLU and maxpooling layers.

  `C64`-BN-ReLU-**ESAB**-RB64-RB128-RB256-RB512

- **Attention PoseNet Decoder**

  $\texttt{C}_r\texttt{256}$ - $\texttt{C}_r\texttt{256}$ - $\texttt{C}_r\texttt{256}$ - `C6`

The overview for Attention PoseNet is given in Fig. 4.1 b and the details of the attention mechanism are introduced in next subsection.

4.3 Endo-SfMLearner Spatial Attention Block (ESAB)

The intuition behind the ESAB module in encoder layers is to guide pose network by emphasizing texture details and depth differences of pixels. On the contrary to feature-based and object-based attentions, spatial attention selects a specific region of the input image and features in that regions are processed by attention block. The ESAB mechanism is non-local convolutional process. For any given input $\mathbf{X} \in \mathcal{R}^{N\times 64\times H\times W}$, our block operation can be overviewed as:

$$\mathbf{Z} = f(\mathbf{X}, \mathbf{X}^\top) g(\mathbf{X}), \tag{4.1}$$

where f stands for the pixelwise relations of input $\mathbf{X}$ between each pixel. The non-local operator extracts the relative weights of all positions on the feature maps.

In ESAB Block, we employ the dot product operation on max-pooled ϕ and θ convolution, which is activated by ReLU function:

$$\mathbf{P} = \psi(\sigma_{relu}(\theta(\mathbf{X})\phi(\mathbf{X})^\top)), \tag{4.2}$$

where σ_{relu} is the ReLU activation function. The dot product, $\theta(\mathbf{X})\varphi(\mathbf{X})^\top$, gives a measurement for the input covariance, which can be defined as a degree of tendency between two feature maps at different channels. We activate the ψ convolution oper-

ation in *softmax* function and perform a matrix multiplication between the g and the output of *softmax* function. Then, we convolve and upsample the result of multiplication with ϕ to extract the attention map $\mathbf{S}$. Finally, an element-wise sum operation in between attention map $\mathbf{S}$ and the input $\mathbf{X}$ generates the output $\mathbf{E} \in \mathbb{R}^{N \times 64 \times H \times W}$:

$$\mathbf{S} = \phi(\sigma_{softmax}(\mathbf{P})g(\mathbf{X})), \tag{4.3}$$

$$\mathbf{F} = \mathbf{S} + \mathbf{X}, \tag{4.4}$$

where $\sigma_{softmax}$ denotes *softmax* function. Short connection between the input $\mathbf{X}$ and the output $\mathbf{F}$ finalizes the block operations for the residual learning. The detailed flow diagram of block operations of ESAB module is given in Fig. 4.1 d.

4.4 Learning Objectives for Endo-SfMLearner

Endo-SfMLearner is trained both in forward and backward directions with losses calculated in forward direction. We are using three loss functions to guide the network without labels; brightness-aware photometric loss, smoothness loss, and geometry consistency loss. Apart from well-known way of defining photometric loss, we are proposing affine brightness transformation between consecutive frames to deal with the problems stem from brightness constancy assumption of previous methods. First of all, the new reference image, $\hat{I}_i$, is synthesized via interpolating I_{i+1}. Previous methods calculate photometric loss directly comparing the synthesized image $\hat{I}_i$ with target image, I_i. However, the difference stem from illumination between consecutive frames might mislead the network. We propose to equate the brightness conditions between these two images as a robust way of supervising training phase. To the best of our knowledge, this is the first implementation of that approach for pose and depth estimation in literature. Moreover, quickly changing the distance between organ tissue and camera results in scale inconsistency. We are using geometry consistency loss [63] to cope with that problem. The overall objective of the system is to minimize the weighted sum of brightness-aware photometric loss $\mathcal{L}_{bp}^{M}$, smoothness loss $\mathcal{L}_s$ and geometry consistency

loss $\mathcal{L}_{GC}$ which can be formulated as:

$$\mathcal{L} = \omega_1 \mathcal{L}_{bp}^{M} + \omega_2 \mathcal{L}_s + \omega_3 \mathcal{L}_{GC}. \tag{4.5}$$

where ω_1, ω_2, and ω_3 are the weights for the related loss functions which are not necessarily adding up to one.

The well-known photometric loss functions are based on the brightness constancy assumption which can be violated due to auto-exposure of the camera and fast illumination changes to which both L_2 and SSIM are no more invariant. To deal with that inconsistent illumination issue which is common in endoscopic image sequences, Endo-SfMLearner network predicts a brightness transformation parameter set which tries to align the brightness of input images during training on the fly and in a self-supervised manner. The evaluations demonstrate that the proposed brightness transformation significantly improves the pose and depth prediction accuracy. The brightness-aware photometric loss formulation is given as follows:

$$\mathcal{L}_{bp} = \frac{1}{|P|} \sum_{p \in P} (\lambda_p \| \mathbf{T}_b(\hat{I}_i(p)) - I_i(p) \|_2 + \lambda_s \frac{1 - SSIM_{\mathbf{T}_b(\hat{I}_i), I_i}}{2}) \tag{4.6}$$

$$\mathbf{T}_b(\hat{I}_i) = \hat{I}_i^{a_{t'}, b_{t'}} = a_{t \to t'} \hat{I}_i + c_{t \to t'} \tag{4.7}$$

where $\hat{I}_i$ stands for the synthesized image by warping I_{i+1}, $\mathbf{T}_b$ is the brightness alignment function with affine transformation parameters $a_{t \to t'}$ and $c_{t \to t'}$, P stands for the successfully projected pixels from reference frame, $SSIM$ is the image dissimilarity loss. By making use of contrast, luminance, and structure values of $\mathbf{T}_b(\hat{I}_i)$ and I_i image; SSIM targets to measure perceived image quality by human visual system and

more sensitive to high frequency content such as textures and edges in regard of PSNR.

Since the photometric loss is not sufficiently informative for the low-texture and homogeneous endoscopic images, we are also incorporating smoothness loss [64] which is calculated as a combination of predicted depth and input images for both reference and target frames.

$$\mathcal{L}_s = \sum_{p \in P} (e^{-\nabla I_i(p)} . \nabla D_i(p))^2, \tag{4.8}$$

where ∇ is the first derivative along spatial directions. Thanks to the smoothness loss, Endo-SfMLearner is guided by edges in the predicted depth and input images. Finally, geometry consistency loss is integrated into our methodology. The main idea behind this loss is to confirm if D_i provides the same scene under the transformation of D_{i+1} by predicted relative poses $P_{i,i+1}$. The difference between predicted depths, D_{diff}, can be calculated as:

$$D_{diff}(p) = \frac{|D^i_{i+1}(p) - D'_{i+1}(p)|}{D^i_{i+1}(p) + D'_{i+1}(p)}, \tag{4.9}$$

where D^i_{i+1} is the depth map of I_{i+1} by warping D_i via $P_{i,i+1}$ and D'_{i+1} is the interpolated depth map from D_{i+1}. The geometry consistency loss will be defined as summation of this difference across all pixel coordinates after normalization with valid pixel counts:

$$\mathcal{L}_{GC} = \frac{1}{|P|} \sum_{p \in P} D_{diff}(p). \tag{4.10}$$

This consistency constrain between consecutive depth maps paves the way for long trajectory estimation with a higher accuracy, the reader is referred to see Fig. 4.1 a. We also use depth inconsistency map results, D_{diff}, to weight the $\mathcal{L}_{bp}$ with M as follows:

$$M = 1 - D_{diff}, \tag{4.11}$$

$$\mathcal{L}^M_{bp} = \frac{1}{|P|} \sum_{p \in P} (M(p).(\mathcal{L}_{bp}(p))). \tag{4.12}$$

Thanks to this operation, brightness-aware photometric loss is weighted with higher constant if the predicted and interpolated depth maps are inconsistent for each pixel.

4.5 Endo-SfMLearner Architecture Overview

First of all, two consecutive unlabeled images (I_i, I_{i+1}) are fed into depth network separately and their corresponding dense disparity maps are predicted (D_i, D_{i+1}). PoseNet outputs the relative 6D camera poses $P_{i,i+1}$ for the same snippet. Reference images, $\hat{I}_i$, are synthesized with predicted depth and pose by warping the source image I_{i+1}. The difference between $\mathbf{T}_b(\hat{I}_i)$ and I_i master the brightness-aware photometric loss. To deal with the violation of geometric assumptions in image reconstruction (due to insufficiency of endoscope cameras), we also use geometry consistency loss which takes into account the difference between warped D^i_{i+1} and interpolated D'_{i+1} pixel-wise disparity estimation.

DispNet encoder share similar structure with PoseNet encoder except ESAB block and skip connections. Decoder consists of five layers each consists of two convolution layers followed by ELU activation function. With the final convolution operation followed by sigmoid activation function, it outputs the dense disparity map from single image.

The non-local operator deduces the relative weights of all positions on the feature maps which measures the input covariance as a degree of tendency between two feature maps at different channels. For GPU memory usage efficiency which is crucial in global attention applications, max-pooling operations are integrated into the block operations. Thanks to the attention mechanism, PoseNet selectively focuses on texture details for more accurate pose and orientation estimation. The codes and the link for the dataset are publicly available at `https://github.com/CapsuleEndoscope/EndoSLAM`

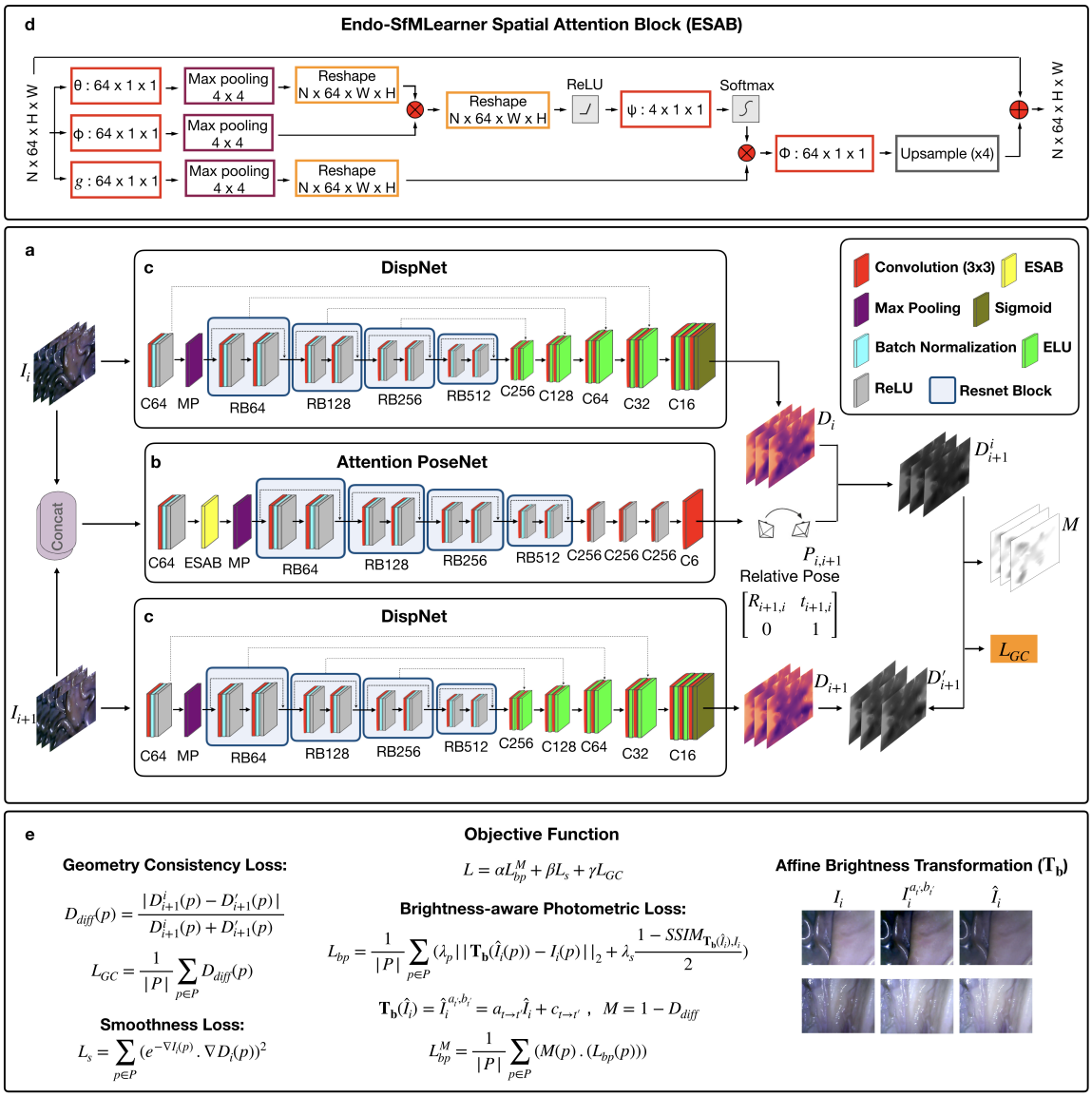


Figure 4.1 Endo-SfMLearner architecture overview. **a** Firstly, two consecutive unlabeled images (I_i, I_{i+1}) are fed into depth network separately and their corresponding dense disparity maps are predicted (D_i, D_{i+1}). PoseNet outputs the relative 6D camera poses $P_{i,i+1}$ for the same snippet. Reference images, $\hat{I}_i$, are synthesized with predicted depth and pose by warping the source image I_{i+1}. The difference between $\mathbf{T}_b(\hat{I}_i)$ and I_i master the brightness-aware photometric loss. To deal with the violation of geometric assumptions, we also use geometry consistency loss which takes into account the difference between warped D^i_{i+1} and interpolated D'_{i+1} pixel-wise disparity estimation. **b** Attention-PoseNet open form. The encoder part of the network consists of four basic ResNet blocks with spatial attention module in between ReLU and maxpooling layer. **c** DispNet encoder share similar structure with PoseNet encoder except ESAB block and skip connections and outputs the dense disparity map from single image. **d** For GPU memory usage efficiency which is crucial in global attention applications, max-pooling operations. Thanks to the attention mechanism, PoseNet selectively focuses on texture details for more accurate pose and orientation estimation. **e** We are using a weighted sum of brightness-aware photometric loss, smoothness loss, and geometry consistency loss as an overall learning objective. Affine brightness transformation function is utilized to equate the illumination conditions in between reference and target image before calculating SSIM and their pixelwise channel differences.

4.6 EndoSLAM Use-Case with Endo-SfMLearner

To illustrate the use-case of the EndoSLAM dataset, Endo-SfMLearner, our proposed learning-based structure-from-motion method was benchmarked for the pose and depth estimation tasks. Additionally, we have tested both dataset and EndoSfMLearner with a traditional fully dense 3D-reconstruction pipeline based on SIFT feature-matching and non-lambertian surface reconstruction where the detailed overview is given in Algorithm 1. Error metrics that were used to quantitatively assess the performance of the algorithms are introduced in the following subsections.

4.6.1 Error Metrics

Endo-SfMLearner pose estimation performance is tested based on three metrics: absolute trajectory error(ATE), translational relative pose error(trans RPE) and rotational relative pose error(rot RPE). The monocular depth estimation performance is evaluated in terms of Root Mean Square Error(RMSE). Finally, the 3D-reconstruction results are evaluated with surface reconstruction error. These error metrics are defined as follows based on the estimated and ground truth trajectories represented by $\mathbf{P}_1, \ldots, \mathbf{P}_n \in \mathrm{SE}(3)$ and $\mathbf{Q}_1, \ldots, \mathbf{Q}_n \in \mathrm{SE}(3)$, respectively, where the lower subscript is indexing frames and SE(3) is the Special Euclidean Group in three dimensions.

4.6.1.1 Absolute trajectory error (ATE). The ATE is a measure of global consistency between two trajectories, comparing absolute distances between ground truth and predicted poses at each point in time. Let the rigid body transformation $\mathbf{S}$ be the best (least-squares) alignment of the trajectories [65]. Then absolute trajectory error for the i^{th} pose sample is calculated as follows:

$$\mathrm{ATE}_i = \|\mathrm{trans}(\mathbf{Q}_i^{-1}\mathbf{S}\mathbf{P}_i)\|. \tag{4.13}$$

The overall error throughout trajectory is defined by the root mean square of ATE_i.

4.6.1.2 Relative Pose Error (RPE). Relative pose error measures the difference in the change in pose over a fixed length Δ between two trajectories. Defining $\mathbf{E}_i(\Delta) = (\mathbf{Q}_i^{-1}\mathbf{Q}_{i+\Delta})^{-1}(\mathbf{P}_i^{-1}\mathbf{P}_{i+\Delta})$, the translational and rotational RPE are given by:

$$\text{Trans RPE}_i(\Delta) = \|\text{trans}(\mathbf{E}_i)\|, \tag{4.14}$$

$$\text{Rot RPE}_i(\Delta) = \angle(\text{rot}(\mathbf{E}_i)), \tag{4.15}$$

where $\text{rot}(\mathbf{E}_i)$ is the rotation matrix of $\mathbf{E}_i$ and $\angle(\cdot)$ is the positive angle of rotation. The errors are reported for Δ equals to 1.

4.6.1.3 Surface Reconstruction Error. We use the methodology propounded by [66] in order to evaluate the surface reconstruction quality. As the first step, one line segment is manually identified between the reconstructed and ground truth 3D maps. The match points are used to coarsely align both maps. This coarse alignment is used as an initialization for the iterative closest point (ICP) algorithm. ICP iteratively aligns both maps until a termination criterion of 0.001 cm deviation in RMSE is reached.

4.6.2 Pose Estimation with Endo-SfMLearner

All methods including Endo-SfMLearner are trained with the same data and parameter set for the sake of fairness and unbiased results. The training and validation dataset consist of 2,039 and 509 colon images generated in the Unity simulation environment, respectively. We train all networks in 200 epochs with randomly shuffled batches each size of 4 images, optimize by ADAM with an initial learning rate 10^{-4} and validate after each epoch. According to the tests in terms of ATE, trans RPE, and rot RPE on the data recorded via the HighCam and LowCam, Endo-SfMLearner achieves the state-of-the-art for most of the cases. The results in Table 4.1 show clear advantage of ESAB block integration and brightness-aware photometric loss. In the majority of Stomach-III results for both HighCam and LowCam, all models fail to follow trajectory with sufficient accuracy. However, the predicted trajectories aligned with ground truth for Endo-SfMLeaner in general are much better compared to other

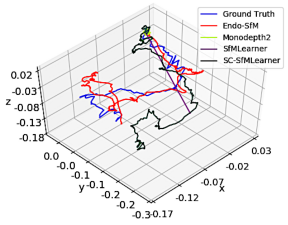
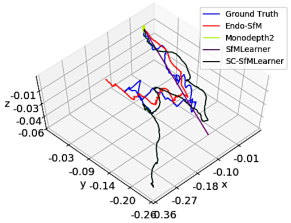
models. Both quantitative and qualitative pose estimation results on sample trajectories of HighCam, LowCam, and MiroCam are given in Fig. 4.2 a-c. Under the above mentioned training conditions, Monodepth2 and SfMLearner face with fatal failure on endoscopic videos. The given results on small intestine illustrates the case where SC-SfMLearner loose the scale consistency after sharp corner angle which is not the case for Endo-SfMLearner. The same problems observed for SfMLearner and Monodepth2 as in the previous case. Endo-SfMLearner tracks loopy sections of the trajectories with sufficient precision up to 1000 frames by leaving a small offset in between the ground truth. Apart from these, camera orientation estimations significantly improve and rotational relative pose error reduces almost three times compared to SC-SfMLearner which is the baseline state-of-the art method and achieves the closest performance to ours. By also exhibiting pose estimation performance of our proposed approach on Mirocam records, we have tested the reliability of Endo-SfMLearner against camera intrinsic properties. Even if SC-SfMLearner exhibits the closest performance to our method in terms of absolute trajectory errors, we observed improvement specially on the rotational movement estimations which is reflected on rotational relative pose errors. A similar observation is also made on Unity trajectories, see Fig. 4.2d-f. On the contrary to real ex-vivo records, synthetically generated trajectories are more straightforward and easier to follow. This fact results in increase in the performance of all methods. However, SC-SfMLeaner and Endo-SfMLeaner track the route with higher accuracy thanks to the geometry consistency loss. Even if all algorithms trained by the synthetic colon images, Monodepth2 and SfMLearner face with the same problem as in real trajectories. For all synthetically generated trajectories, EndoSfMLearner exhibits lowest mean absolute trajectory error(ATE). Although quantitatively Monodepth2 and SfMLearner have lower rotational error, it cannot be taken into account as performance superiority. Since the rotations cannot be changed frequently and easily while recording clear images in Unity environment, they remain close to identity matrix which is generally predicted by Monodepth2 and SfMLearner. Throughout the all ex-vivo trajectory, SfMLearner relative pose estimations vary incredibly small which results in almost straightforward global pose estimation. In the same cases, we observed network firing problem for Monodepth2 even if we have repeated the tests on dataset with different frame-per-second rates. However, SC-SfMLearner and Endo-SfMLearner

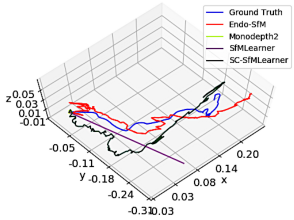
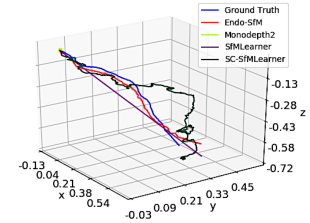
exhibit more reasonable predictions thanks to the geometry consistency loss. The most challenging part of trajectories are sharp corners where the position and orientation of camera change with high speed in small time intervals. At those points, Endo-SfMLearner yields performance with higher accuracy compared to SC-SfMLearner not just qualitatively but also quantitatively in terms of both translational, and rotational errors. Since the rotations cannot be changed frequently and easily while recording clear images in the Unity environment, the trajectories are close to the straight lines which result in higher accuracy for all methods. It is seen that the Endo-SfMLearner outputs generally follow the shape of the ground truth, specifically, it catches rotations more consistently which is the main reason for the decrease in rotational relative pose error.

For more comprehensive evaluations of results in terms of camera motions, descriptive analysis of the camera speeds and accelerations are given in Fig. 3.9, Table 3.6 and Table 3.7. Since the robot motions are highly effective on image quality, we expect a decrease in the pose estimation accuracy for the trajectories of Stomach-III which have highest mean speed and acceleration. The fact paves the way for the difficulty in the alignment of those trajectories and also stitching of those frames for 3D reconstruction.

Table 4.1
Quantitative results of pose prediction for various organs and trajectories.
Endo-SfMLearner comparison with Endo-SfMLearner without attention block(Ew/oAtt), Endo-SfMLearner without brightness aware photometric loss integration(Ew/oBr), SC-SfMLearner, Monodepth2, and SfMLearner. To test the algorithm robustness against tissue and trajectory differences, we performed tests on two separate trajectories from ex-vivo porcine stomach, colon, and intestine. Absolute trajectory error (ATE) is used to quantify the overall consistency throughout path, instead Translational and rotational Relative Pose Error are local metrics. Moreover, for a better understanding of the camera specifications' effect on pose estimation, we compared the results from high (HighCam) and low (LowCam) resolution camera for same trajectories. We observed a considerable decrease in rotational errors for Endo-SfMLearner with respect to the baseline method, SC-SfMLearner which proves the effectiveness of spatial attention block integrated to pose network encoder and brightness-aware photometric loss. Even though, most of the tests result in Endo-SfMLearner superiority, only the third trajectory of Stomach-III from HighCam SC-SfMLearner performed with higher accuracy in terms of ATE. Nevertheless, ablation studies do not provide sufficient cue to explain this improvement either stem from SAB or brightness aware photometric loss.

	Organ, Trajectory	Trajectory Length [m]	ATE ↓ (mean± std) [m]	Trans. RPE ↓ (mean± std) [m]	Rot. RPE ↓ (mean± std) [deg]	Trajectory Length [m]	ATE ↓ (mean± std) [m]	Trans RPE ↓ (mean± std) [m]	Rot RPE ↓ (mean± std) [deg]
		HighCam				LowCam			
EndoSfM	Colon-IV,Traj-I	0.4286	0.0878± 0.0549	0.0009± 0.0027	0.488± 0.3217	0.6785	0.1046± 0.0343	0.0011± 0.006	0.4666± 1.3792
	Colon-IV,Traj-V	1.2547	0.1731± 0.1179	0.0014± 0.002	0.2552± 0.417	1.1699	0.1771± 0.1177	0.0012± 0.002	0.1493± 0.2321
	Intestine,Traj-IV	1.0557	0.0812± 0.0152	0.0010 ± 0.0013	0.173± 0.1942	0.8265	0.0558± 0.0356	0.0011± 0.0008	0.404 ± 0.5052
	Stomach-I,Traj-I	1.4344	0.1183± 0.1062	0.0013± 0.0028	0.5988± 0.8185	0.8406	0.1732± 0.116	0.0021± 0.0034	0.8424± 1.0788
	Stomach-III,Traj-III	0.8908	0.1177± 0.0543	0.0013± 0.0033	0.5543± 0.928	0.9714	0.1014± 0.0491	0.0011± 0.0007	0.6705± 0.3817
Ew/oAtt	Colon-IV,Traj-I	0.4286	0.0894± 0.0274	0.0010± 0.0029	0.3502± 0.2621	0.6785	0.1548± 0.0591	0.0010± 0.3679	1.3613± 1.5908
	Colon-IV,Traj-V	1.2547	0.1855 ± 0.0494	0.0014± 0.0022	0.4569± 0.5734	1.1699	0.1628± 0.0375	0.0014± 0.003	0.4168± 0.3149
	Intestine,Traj-IV	1.0557	0.1055± 0.0379	0.0011± 0.0012	0.3343± 0.2653	0.8265	0.0691± 0.0305	0.001± 0.0009	0.654 ± 0.6042
	Stomach-I,Traj-I	1.4344	0.1889 ± 0.0497	0.0015± 0.0038	0.893± 0.915	0.8406	0.1968± 0.1417	0.0025± 0.0037	1.1823± 1.2112
	Stomach-III,Traj-III	0.8908	0.1362 ± 0.068	0.0016± 0.0032	0.8244 ± 1.0127	0.9714	0.1204± 0.0418	0.0010± 0.0009	1.0907± 0.5634
Ew/oBr	Colon-IV,Traj-I	0.4286	0.1328± 0.0431	0.0010± 0.0026	0.7198± 0.4764	0.6785	0.1402 ± 0.0671	0.0010± 0.0060	0.7257± 1.424
	Colon-IV,Traj-V	1.2547	0.1898± 0.0709	0.0015± 0.002	0.929± 0.7525	1.1699	0.1503± 0.0433	0.0013± 0.002	0.8989± 0.6199
	Intestine,Traj-IV	1.0557	0.1467± 0.0848	0.002 ± 0.0010	0.6607± 0.3884	0.8265	0.1241 ± 0.0436	0.0009± 0.0008	1.106 ± 0.8081
	Stomach-I,Traj-I	1.4344	0.1963 ± 0.0478	0.002 ± 0.0032	0.6899 ± 1.0401	0.8406	0.1923± 0.118	0.0023± 0.0032	0.9215± 1.1728
	Stomach-III,Traj-III	0.8908	0.1277± 0.0805	0.0014± 0.0033	0.3933± 0.9258	0.9714	0.1101 ± 0.0257	0.0010± 0.0006	0.439 ± 0.2672
SC-SfM	Colon-IV,Traj-I	0.4286	0.1545± 0.0441	0.0014± 0.0028	1.3532± 0.8541	0.6785	0.1898± 0.0718	0.0015± 0.0060	1.6388± 1.5908
	Colon-IV,Traj-V	1.2547	0.2054± 0.1734	0.0024± 0.0029	1.2452± 0.965	1.1699	0.1667± 0.1263	0.0021± 0.003	1.2188± 0.7715
	Intestine,Traj-IV	1.0557	0.1247± 0.1327	0.0015± 0.0009	0.9257± 0.584	0.8265	0.0908± 0.0819	0.0016± 0.0009	0.8989 ± 0.7854
	Stomach-I,Traj-I	1.4344	0.2325± 0.127	0.002± 0.0038	1.2937± 1.2484	0.8406	0.191± 0.1399	0.0028± 0.0033	2.1322± 1.2601
	Stomach-III,Traj-III	0.8908	0.0898 ± 0.035	0.0016± 0.0033	1.3071± 1.3187	0.9714	0.1927± 0.0561	0.0012± 0.0007	2.041± 0.8391
Mono2	Colon-IV,Traj-I	0.4286	0.1071± 0.0756	0.0012± 0.0028	0.3115± 0.268	0.6785	0.215± 0.1084	0.0009± 0.006	0.1679± 1.378
	Colon-IV,Traj-V	1.2547	0.1872± 0.1404	0.0016± 0.002	0.1607± 0.4226	1.1699	0.2158± 0.1466	0.0018± 0.002	0.3921± 0.3362
	Intestine,Traj-IV	1.0557	0.1507± 0.1165	0.009 ± 0.0013	0.1092± 0.1812	0.8265	0.1431± 0.132	0.0014± 0.001	0.3128 ± 0.5288
	Stomach-I,Traj-I	1.4344	0.2878 ± 0.2293	0.0029± 0.0038	0.298± 0.7968	0.8406	0.2033± 0.0971	0.0019± 0.0011	0.5296± 0.3642
	Stomach-III,Traj-III	0.8908	0.5841± 0.2742	0.0022± 0.0033	0.8178± 0.9059	0.9714	0.3876± 0.2322	0.0032± 0.0017	0.7345± 0.8349
SfM	Colon-IV,Traj-I	0.4286	0.1584±0.1064	0.0043± 0.0042	2.6624± 1.6822	0.6785	0.1946± 0.1708	0.0037± 0.0092	2.0718± 2.3018
	Colon-IV,Traj-V	1.2547	0.5849± 0.5201	0.0092± 0.0175	4.4083± 4.6309	1.1699	0.2094± 0.1613	0.005± 0.0041	3.1999± 1.8304
	Intestine,Traj-IV	1.0557	0.2119± 0.2022	0.0083± 0.016	3.9877± 5.2134	0.8265	0.2387± 0.1675	0.0048± 0.005	2.7019± 2.189
	Stomach-I,Traj-I	1.4344	0.1741± 0.0744	0.0012± 0.0038	0.7249± 0.7904	0.8406	0.2226± 0.0989	0.007± 0.005	4.1709± 2.3479
	Stomach-III,Traj-III	0.8908	0.3086± 0.1774	0.0018± 0.0035	0.6137± 0.996	0.9714	0.1711± 0.0548	0.0012± 0.0008	0.802± 0.4236

	a Small Intestine - LowCam				b Small Intestine - HighCam			
Trajectory Plots								
Number of Frames, FPS	1190, 20				1524, 20			
Trajectory Length [m]	0.8459				1.2027			
Algorithms	Endo-SfM	SC-SfMLearner	Monodepth2	SfMLearner	Endo-SfM	SC-SfMLearner	Monodepth2	SfMLearner
ATE (mean±std) [m]	0.201±0.1721	0.2302±0.1167	0.436±0.1860	0.6247±0.5883	0.0481±0.212	0.0549±0.0173	0.117±0.0412	0.0623±0.0236
Trans. RPE (mean±std) [m]	0.0013±0.0009	0.0018±0.0008	0.0025±0.0012	0.0155±0.0426	0.0009±0.0007	0.0012±0.0009	0.0014±0.0009	0.0008±0.0007
Rot. RPE (mean±std) [deg]	0.4215±0.2767	1.0348±0.6951	0.0023±0.0009	4.5832±3.8278	0.5156±0.3996	1.5035±0.9861	0.2171±0.3159	0.2166±0.3162

	c Small Intestine - MiroCam				d Colon - UnityCam			
Trajectory Plots								
Number of Frames, FPS	137, 3				900, 30			
Trajectory Length [m]	0.6361				0.9665			
Algorithms	Endo-SfM	SC-SfMLearner	Monodepth2	SfMLearner	Endo-SfM	SC-SfMLearner	Monodepth2	SfMLearner
ATE (mean±std) [m]	0.2476±0.1673	0.3617±0.1498	0.3212±0.1232	0.4183±0.2427	0.2468±0.1137	0.3582±0.1717	0.4053±0.2197	0.6417±0.1711
Trans. RPE (mean±std) [m]	0.0053±0.0037	0.0069±0.0048	0.0044±0.0017	0.0043±0.001	0.0013±0.0007	0.0034±0.0029	0.0025±0.0012	0.009±0.003
Rot. RPE (mean±std) [deg]	0.5809±0.3864	1.1736±0.6279	0.7717±0.3817	0.4188±0.0236	0.2548±0.2073	0.5586±0.706	0.0067±0.0037	0.0065±0.014

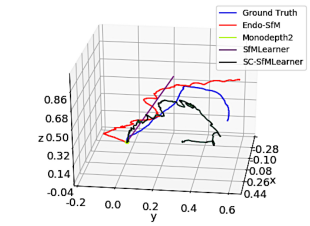
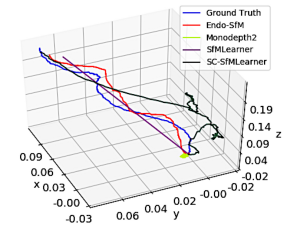
	e Small Intestine - UnityCam				f Stomach - UnityCam			
Trajectory Plots								
Number of Frames, FPS	1257, 30				300, 30			
Trajectory Length [m]	1.3663				0.3267			
Algorithms	Endo-SfM	SC-SfMLearner	Monodepth2	SfMLearner	Endo-SfM	SC-SfMLearner	Monodepth2	SfMLearner
ATE (mean±std) [m]	0.2248±0.0705	0.3142±0.2363	0.5224±0.2629	0.314±0.1494	0.0505±0.0306	0.063±0.0334	0.1337±0.0753	0.1995±0.056
Trans. RPE (mean±std) [m]	0.0019±0.0012	0.0024±0.002	0.0034±0.0016	0.0015±0.0003	0.0011±0.0005	0.0021±0.0011	0.0016±0.0007	0.0011±0.003
Rot. RPE (mean±std) [deg]	0.2278±0.2076	0.4235±0.5359	0.0192±0.0688	0.0183±0.0701	0.2814±0.1496	0.7996±0.6484	0.0073±0.0074	0.0083±0.021

Figure 4.2 Pose estimations. Endo-SfMLearner, SC-SfMLearner, Monodepth2, and SfMLearner trajectory estimations are benchmarked on ex-vivo EndoSLAM data. **a** The results for the first trajectory of small intestine recorded by LowCam. **b** The results for first sub-trajectory of small intestine recorded by HighCam. **c** The results for the fourth trajectory of Colon-III recorded by MiroCam. On the contrary to the HighCam and LowCam, MiroCam exhibits fish-eye camera properties with high lens distortion. Due to more straightforward and easier to follow trajectories the performance increase for all methods. Although quantitatively Monodepth2 and SfMLearner have lower rotational error, it cannot be taken into account as performance superiority. Since the rotations cannot be changed frequently and easily while recording clear images in Unity environment, they remain close to identity matrix which is generally predicted by Monodepth2 and SfMLearner.

4.6.3 Depth Estimation with Endo-SfMLearner

We have performed the tests for the pixel-wise depth estimation quantitatively on the EndoSLAM dataset as well as qualitatively on EndoSLAM, Kvasir [8], and Nerthus dataset [67] with the aim of proving the applicability of the propounded approach in a real endoscopy procedure.

4.6.3.1 Quantitative Evaluation. Since EndoSLAM dataset also provides pixelwise depth ground truth for synthetically generated endoscopic frames, we show that Endo-SfMLearner quantitatively outperforms the benchmarked monocular depth estimation methods as given in Fig. 4.3. The results are evaluated in terms of root mean square error (RMSE) on 1,548 stomach, 1,257 small intestine, and 1,062 colon frames. Even if the training and validation dataset consist of synthetic colon frames, Endo-SfMLearner depicts high performance on stomach and small intestine with 0.2966 and 0.1785 mean RMSE. The heatmaps are also indicating that the errors significantly decrease for the pixels representing regions far from 14mm.

4.6.3.2 Qualitative Evaluation. We employed depth map evaluations qualitatively on the Kvasir, Nerthus and EndoSLAM dataset. Although the ex-vivo part of our dataset does not provide pixel-wise depth ground truth, the results for small intestine trajectory given in Fig. 4.4 d, e exhibits that Endo-SfMLearner is more capable of catching depth alterations even in the short ranges with the support of spatial attention mechanism. Even if all models trained with synthetically generated normal colon frames, the performance on Kvasir normal dataset depicts the data adaptability of the proposed approach. We have also examined the model performance under various texture details on the Kvasir polyps dataset since the polyp regions differ from real tissue not only in terms of shape but also the texture. Fig. 4.4 b indicates that the method successfully detects the boundaries and details of polyps. Moreover, for most of the cases the propounded method generates more consistent depth maps for the pixels where the light reflections occur thanks to the robustness against illumination

changes provided by brightness-aware photometric loss. Especially taking a closer look at the light reflections in the 4.4 e; it can be observed that depth estimation accuracy of SC-SfMLearner at reflective regions decreases drastically which is not the case for EndoSfMLearner. Apart from these, the robustness against blur and radial distortion effects are examined in three scenarios. In Fig. 4.4 d, it can be observed that the center of frames are predicted as closer to the camera with an increase of distortion level. On the contrary, the general tendency for all models is to make longer distance predictions with the increase of the Gaussian Blur effect.

4.6.3.3 Ablation Studies for Spatial Attention Block. In order to increase the pose and depth network sensitivity for the edge and texture details, we have integrated an attention block in between ReLU and max pooling operations in PoseNet encoder. By this attention mechanism, we are expecting to preserve low and high-frequency information from the input endoscopic images by exploiting the feature-channel inter-dependencies. In this subsection, we specifically investigate the following cases:

- EndoSfMLearner with brightness-aware photometric loss and ESAB,
- EndoSfMLearner with ESAB and without brightness-aware photometric loss,
- EndoSfMLearner without ESAB and with brightness-aware photometric loss.

The results for the pose tracking given in Table 4.1 reveal the usefulness and effectiveness of the module. Although the attention module is only inserted in PoseNet, simultaneous training of networks causes the improvement in depth estimation which is depicted in Fig. 4.3. As seen from quantitative ablation analysis, the attention module makes Endo-SfMLearner more responsive for depth alterations on the synthetically generated images of all organs. Even for the stomach and small intestine that is not included in the training phase, Endo-SfMLearner achieves acceptable RMSE values which is the indicator of its persistent effort to be adaptable for texture differences.

Algorithm 1: 3D Reconstruction and Evaluation Pipeline

begin

[1] Extract SIFT features between image pairs

[2] Find k-nearest neighbours for each feature using a k-d tree

for *each image* **do**

(i) Select m candidate matching images that have the most number of corresponding feature points

(ii) Find geometrically consistent feature matches using RANSAC to solve for the homography between pairs of images.

end

[3] Find connected components of image matches

for *each connected component* **do**

(i) Perform bundle adjustment for connected components in image matches

(ii) Render final stitched image using multi-band blending

end

[4] Apply inpainting on the stitched image to suppress specularities

[5] Reconstruct the surface using Tsai-Shah shape from shading method

[6] Label a common line segment in ground truth data and reconstructed surface

[7] Apply ICP algorithm using the common line as initialization

[8] Compute iteratively the cloud-to-mesh distances to acquire RMSE

end

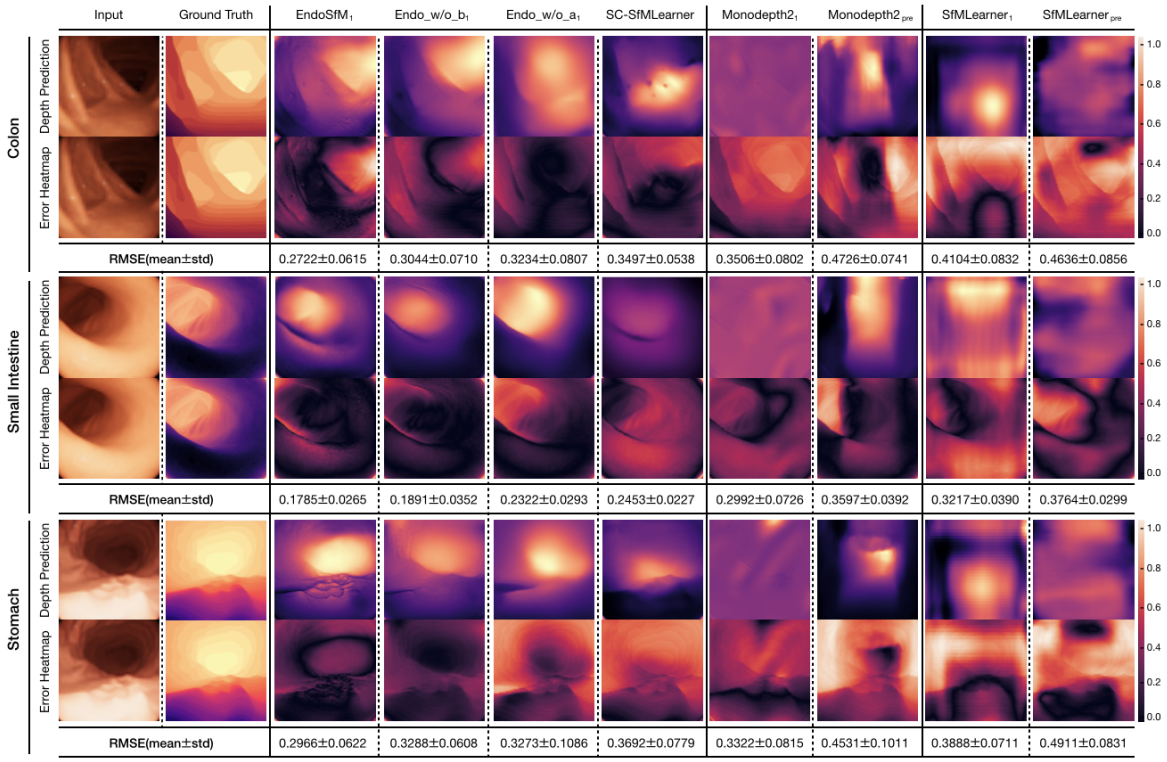


Figure 4.3 Quantitative depth evaluations. The original input image, depth ground truth, predicted depth maps and error heatmaps by Endo-SfMLearner, Endo-SfMLearner without brightness loss integration(Endo_w/o_b_1), Endo-SfMLeaner without attention integration(Endo_w/o_a_1), SC-SfMLearner(Endo-SfMLearner without loss and block operation), Monodepth2, published pretrained Monodepth2(Monodepth2$_{pre}$), SfMLearner and published pretrained SfMLearner(SfMLearner$_{pre}$) are shown from left to right, respectively. We benchmark the algorithms quantitatively on the synthetically generated images acquired with the camera whose properties are equivalent to the MiroCam. Even if the models subscripted by "1" are trained with the same data and parameter set, Endo-SfMLearner and SC-SfMLearner which are guided by geometry consistency loss show considerably superior performance to the rest of the methods. In particular, Endo-SfMLearner is able to estimate the relatively far regions more accurately than the remaining ones, although it is optimized for the images obtained by shallow Depth of Field cameras. Besides, its predictions conform with camera light burst and small depth alterations which result in least RMSE errors for all organs that is also proving the cross-organ adaptability of the method. By comparing the Endo-SfMLearner, Endo_w/o_b_1 and Endo_w/o_a_1, one can deduce that the biggest advantage of ESAB block in PoseNet provided to the DispNet is increasing texture awareness whereas brightness-aware photometric loss focusese the network to the light variations throughout the pixels. Their collaboration significantly improves the performance which is supported by decreasing RMSE values. The published pre-trained models are trained with Kitty dataset generally consist of images whose upper part representing distant sky points, right and left edges are closer points representing flats or moving cars. This fact causes biased depth estimation especially for Monodepth2$_{pre}$, on endoscopic images from all organs.

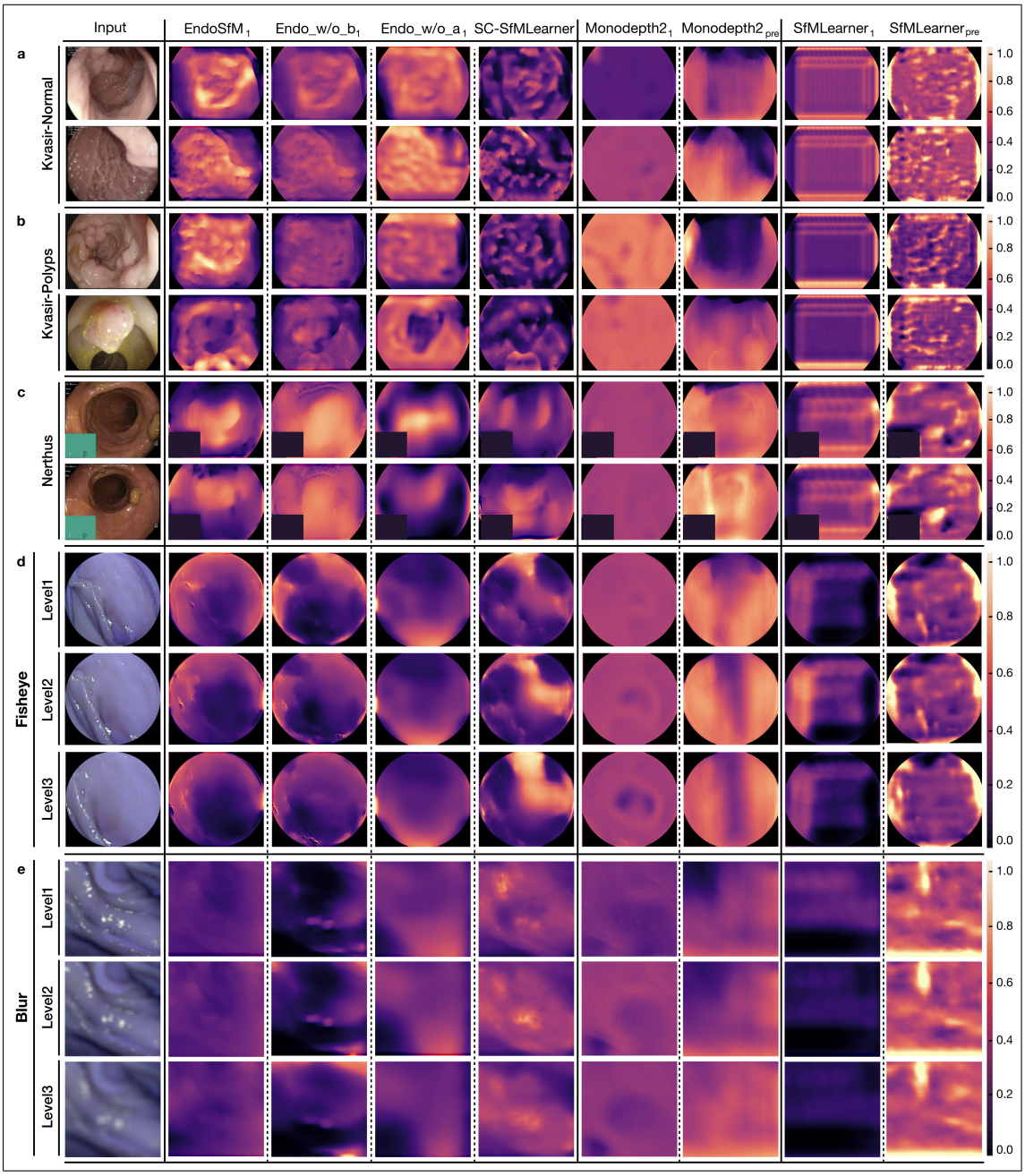


Figure 4.4 Qualitative depth evaluations. The original input image, and predicted depth maps are given for Endo-SfmLearner, Endo-SfMLearner without brightness loss integration(Endo_w/o_b$_1$), Endo-SfMLeaner without attention integration(Endo_w/o_a$_1$), SC-SfMLearner(Endo-SfMLearner without loss and block operation), Monodepth2, published pretrained Monodepth2(Monodepth2$_{pre}$), SfMLearner and published pretrained SfMLearner(SfMLearner$_{pre}$) are shown from left to right, respectively. We benchmark the algorithms qualitatively on **a** the Kvasir normal colon mucosa **b** Kvasir polyps **c**, Nerthus, and **d, e** EndoSLAM dataset. Since the polyp regions differ from real tissue not only in terms of shape but also the texture, we have specially examined the model performance under various texture details on Kvasir polyps dataset, and as seen the polyp boundaries are successfully detected. To illustrate the use-case of data augmentation functions, we have shown that the depth estimation performance on three different radial distortion constant for fish-eye function, as well as, three group under the effect of various Gaussian Blur parameter set. Despite the deficits of the frames, Endo-SfMLearner is capable to cope with the various camera specs. Ablation studies clarify that the attention block provides the awareness for the edges and texture details and brightness aware loss increases the sensitivity of depth estimation for illumination changes. The combined effect of these two achieves the best performance for all cases.

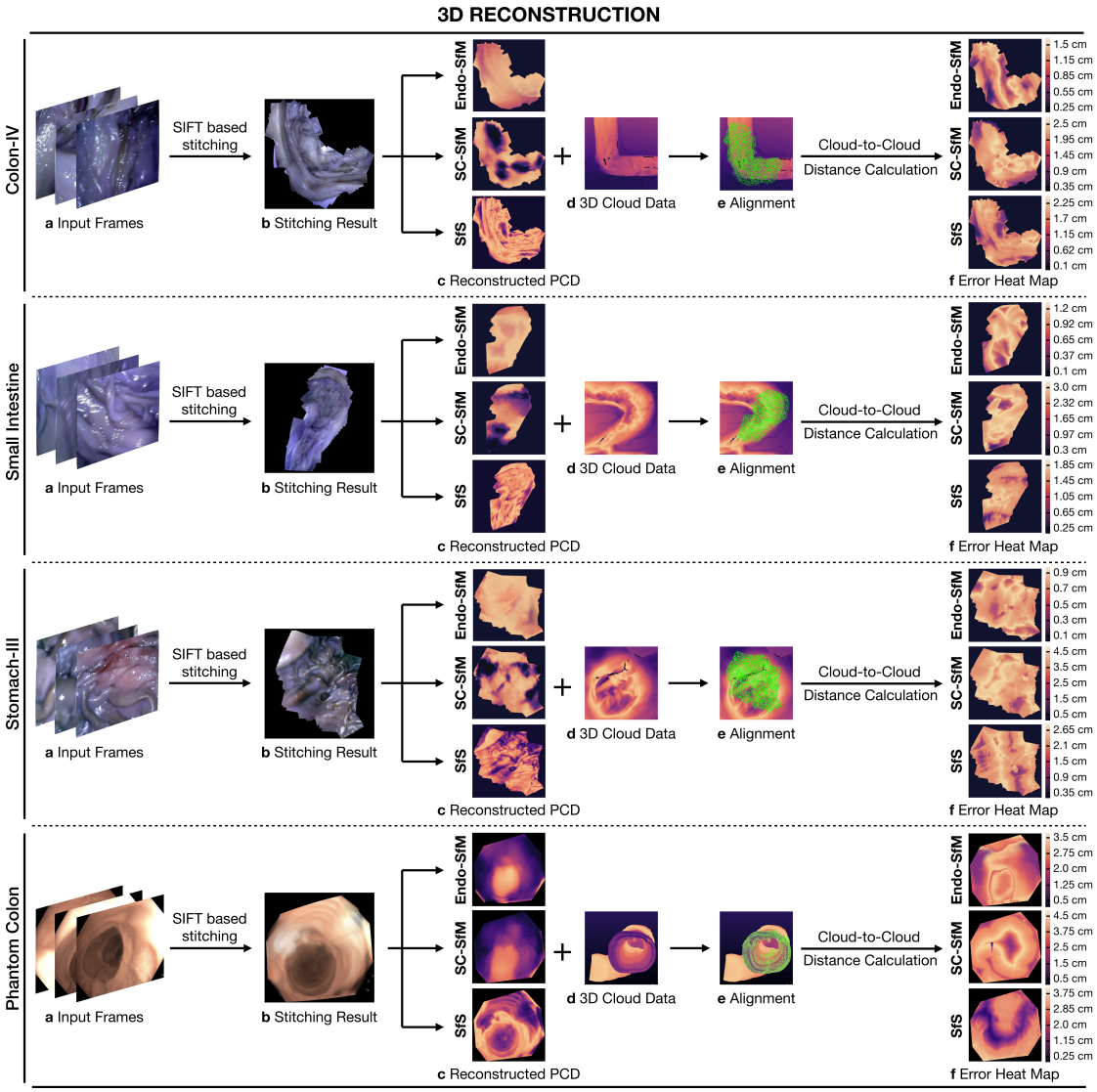


Figure 4.5 3D-Map reconstruction and evaluation pipeline. **a** Input image sequences from Colon-IV, Small Intestine, Stomach-III, and Phantom Colon trajectories which are downsampled to 4 fps. The frames are given as input to Scale Invariant Feature Transform (SIFT), separately. **b** The final stitched image which is formed by aligning and blending all input frames. Specularities are suppressed using the inpainting function of OpenCV. **c** Depth maps for inpainted images which are predicted using Endo-SfMLearner, SC-SfMLearner, and shape from shading. **d** 3D scanner point cloud data for each organ in ply-format. **e** The matched area between reference and aligned cloud points by emphasizing in green colour. The aligned regions are chosen as the same for all compared groups for the sake of fairness. Iterative Closest Point(ICP) was used to align the ground truth data and reconstructed surface after manually labeling a common line segment. **f** The cloud mesh distances in the form of heatmap with the bar displaying the root mean square error in cm. The RMSE values of Colon-IV, 0.51 cm, 0.86 cm, and 0.65 cm for Endo-SfMLearner, SC-SfMLearner, and shape from shading, respectively. The RMSE values of Small Intestine are 0.40 cm, 1.02 cm, and 0.54 cm for Endo-SfMLearner, SC-SfMLearner, and shape from shading, respectively. The RMSE values of Stomach-III are 0.41 cm, 1.37 cm, and 0.73 cm for Endo-SfMLearner, SC-SfMLearner, and shape from shading, respectively. The RMSE values of Phantom Colon are 1.23 cm, 1.56 cm, and 1.38 for Endo-SfMLearner, SC-SfMLearner and shape from shading, respectively. For all organs, we sight the superiority of the Endo-SfMLearner over both SC-SfMLearner and shape from shading. Since the training and validation dataset of SC-SfMLearner consist of colon frames, the RMSE values for colon are smaller than the other organs. However, even if the Endo-SfMLearner has the same training and validation dataset, it exhibits highly effective performance on stitched stomach and intestine images in comparison with the remaining methods.

5. CONCLUSION

In this thesis, we introduce a novel endoscopic SLAM dataset that contains both capsule and standard endoscope camera images with 6D ground truth pose and high precision scanned 3D maps of the explored GI organs. Four different cameras were employed in total to collect data from eight ex-vivo porcine GI-tract organs each from different animal instances. Besides, the dataset also provides an opportunity for the verification of 3D reconstruction algorithms via the recording of the fully covered colon by clinically in use conventional Olympus camera with CT ground truth. Various additional post processing effects such as fisheye distortions, Gaussian blur, downsampling, and vignetting can be applied as optional to diversify and enrich the dataset. In addition to the EndoSLAM dataset, Endo-SfMLearner is proposed as a monocular pose and depth estimation method based on spatial attention mechanisms and brightness-aware hybrid loss. Although Endo-SfMLearner is specifically developed and optimized for the endoscopic type of images, it also holds great promise for laparoscopy images due to similar texture characteristics. Our future work will focus on generalizing the EndoSLAM dataset concept to other visualization techniques and create datasets with various other imaging modalities. Furthermore, we aim to examine and improve the data adaptability of the Endo-SfMLearner and address these issues as next steps. Last but not least, we plan to investigate the combination of Endo-SfmLearner with segmentation, abnormality detection, and classification tasks in the concept of multi-task and meta-learning to enhance the performance of state-of-the-art methods.

5.1 List of publications produced from the thesis

1. EndoSLAM dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic video Kutsev Bengisu Ozyoruk, Guliz Irem Gokceler, Taylor L. Bobrow, Gulfize Coskun, Kagan Incetan, Yasin Almalioglu, Faisal Mahmood, Eva Curto, Luis Perdigoto, Marina Oliveira, Hasan Sahin, Helder Araujo, Henrique Alexandrino, Nicholas J. Durr, Hunter B. Gilbert, Mehmet Turan, *Medical Image Analysis, Volume 71, 2021, 102058, ISSN 1361-8415, https://doi.org/10.1016/j.media.2021.102058.*
(https://www.sciencedirect.com/science/article/pii/S1361841521001043)

REFERENCES

1. Zhang, Z., "Flexible camera calibration by viewing a plane from unknown orientations," in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, Vol. 1, pp. 666–673, Ieee, 1999.

2. American Cancer Society, *Global Cancer Facts and Figures 4th Edition*, American Cancer Society, 2018.

3. Arnold, M., *et al.*, "Global burden of 5 major types of gastrointestinal cancer," *Gastroenterology*, April 2020.

4. Redondo-Cerezo, E., A. Sanchez-Capilla, P. Torre-Rubio, and J. Teresa, "Wireless capsule endoscopy: Perspectives beyond gastrointestinal bleeding," *World journal of gastroenterology : WJG*, Vol. 20, pp. 15664–15673, 11 2014.

5. "An astounding 19 million colonoscopies are performed annually in the united states." `https://idataresearch.com/an-astounding-19-million-colonoscopies-are-performed-annually-in-the-united-states/`, year = 2018, note = Accessed: 08/08/2018.

6. Yano, T., and H. Yamamoto, "Vascular, polypoid, and other lesions of the small bowel," *Best practice & research. Clinical gastroenterology*, Vol. 23, pp. 61–74, 02 2009.

7. Spyrou, E., and D. K. Iakovidis, "Video-based measurements for wireless capsule endoscope tracking," *Measurement Science and Technology*, Vol. 25, no. 1, p. 015002, 2013.

8. Pogorelov, K., *et al.*, "Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection," in *Proceedings of the 8th ACM on Multimedia Systems Conference*, pp. 164–169, 06 2017.

9. Borgli, H., *et al.*, "Hyper-kvasir: A comprehensive multi-class image and video dataset for gastrointestinal endoscopy," 12 2019.

10. Jha, D., P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. D. Johansen, "The Kvasir-SEG Dataset." `https://datasets.simula.no/kvasir-seg/`, 2020.

11. Koulaouzidis, A., *et al.*, "KID Project: an internet-based digital video atlas of capsule endoscopy for research purposes," *Endosc Int Open*, Vol. 5, pp. E477–E483, Jun 2017.

12. Moccia, S., *et al.*, "Learning-based classification of informative laryngoscopic frames," *Computer Methods and Programs in Biomedicine*, Vol. 158, 05 2018.

13. Penza, V., A. S. Ciullo, S. Moccia, L. S. Mattos, and E. De Momi, "Endoabs dataset: Endoscopic abdominal stereo image dataset for benchmarking 3d stereo reconstruction algorithms," *The International Journal of Medical Robotics and Computer Assisted Surgery*, Vol. 14, no. 5, p. e1926, 2018.

14. Bernal, J., J. Sanchez, and F. Vilarino, "Towards automatic polyp detection with a polyp appearance model," *Pattern Recognition*, Vol. 45, pp. 3166–3182, 09 2012.

15. Bernal, J., *et al.*, "Comparative validation of polyp detection methods in video colonoscopy: Results from the miccai 2015 endoscopic vision challenge," *IEEE Transactions on Medical Imaging*, Vol. 36, pp. 1231–1249, June 2017.

16. Bernal, J., *et al.*, "Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Computerized Medical Imaging and Graphics*, Vol. 43, p. 99â111, July 2015.

17. Silva, J. S., A. Histace, O. Romain, X. Dray, and B. Granado, "Towards embedded detection of polyps in WCE images for early diagnosis of colorectal cancer," *International Journal of Computer Assisted Radiology and Surgery*, Vol. 9, no. 2, pp. 283–293, 2014.

18. Tajbakhsh, N., S. R. Gurudu, and J. Liang, "Automated polyp detection in colonoscopy videos using shape and context information," *IEEE Transactions on Medical Imaging*, Vol. 35, pp. 630–644, Feb 2016.

19. Ye, M., S. Giannarou, A. Meining, and G.-Z. Yang, "Online tracking and retargeting with applications to optical biopsy in gastrointestinal endoscopic examinations," *Medical image analysis*, Vol. 30, pp. 144–157, 2016.

20. Ye, M., E. Johns, A. Handa, L. Zhang, P. Pratt, and G.-Z. Yang, "Self-supervised siamese learning on stereo image pairs for depth estimation in robotic surgery," *arXiv preprint arXiv:1705.08260*, 2017.

21. "Robust medical instrument segmentation (robust-mis) challenge 2019." `https://www.synapse.org/\#!Synapse:syn18779624/wiki/592660`. Accessed: 2020-02-12.

22. Dey, N., A. S. Ashour, F. Shi, and R. S. Sherratt, "Wireless capsule gastrointestinal endoscopy: Direction-of-arrival estimation based localization survey," *IEEE reviews in biomedical engineering*, Vol. 10, pp. 2–11, 2017.

23. Shah, T., S. M. Aziz, and T. Vaithianathan, "Development of a tracking algorithm for an in-vivo rf capsule prototype," in *2006 International Conference on Electrical and Computer Engineering*, pp. 173–176, IEEE, 2006.

24. Son, D., S. Yim, and M. Sitti, "A 5-d localization method for a magnetically manipulated untethered robot using a 2-d array of hall-effect sensors," *IEEE/ASME Transactions on Mechatronics*, Vol. 21, no. 2, pp. 708–716, 2015.

25. Kuth, R., J. Reinschke, and R. Rockelein, "Method for determining the position and orientation of an endoscopy capsule guided through an examination object by using a navigating magnetic field generated by means of a navigation device," Feb. 15 2007. US Patent App. 11/481,935.

26. Than, T. D., *et al.*, "An effective localization method for robotic endoscopic capsules using multiple positron emission markers," *IEEE Transactions on Robotics*, Vol. 30, no. 5, pp. 1174–1186, 2014.

27. Ciuti, G., *et al.*, "Frontiers of robotic endoscopic capsules: a review," *Journal of Micro-Bio Robotics*, Vol. 11, no. 1-4, pp. 1–18, 2016.

28. Simaan, N., R. H. Taylor, and H. Choset, "Intelligent surgical robots with situational awareness," *Mechanical Engineering*, Vol. 137, no. 09, pp. S3–S6, 2015.

29. Turan, M., Y. Almalioglu, H. Araújo, E. Konukoglu, and M. Sitti, "Deep endovo: A recurrent convolutional neural network (RCNN) based visual odometry approach for endoscopic capsule robots," *CoRR*, Vol. abs/1708.06822, 2017.

30. Hartley, R., and A. Zisserman, *Multiple View Geometry in Computer Vision*, New York, NY, USA: Cambridge University Press, 2 ed., 2003.

31. Wu, C., S. Agarwal, B. Curless, and S. M. Seitz, "Multicore bundle adjustment," in *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR*, pp. 3057–3064, IEEE, 2011.

32. Leonard, S., A. Sinha, A. Reiter, M. Ishii, G. Gallia, R. Taylor, and G. Hager, "Evaluation and stability analysis of video-based navigation system for functional endoscopic sinus surgery on in-vivo clinical data," *IEEE Transactions on Medical Imaging*, May 2018.

33. Grasa, O. G., E. Bernal, S. Casado, I. Gil, and J. Montiel, "Visual slam for handheld monocular endoscope," *IEEE transactions on medical imaging*, Vol. 33, no. 1, pp. 135–146, 2013.

34. Liu, X., A. Sinha, M. Ishii, G. D. Hager, A. Reiter, R. H. Taylor, and M. Unberath, "Dense depth estimation in monocular endoscopy with self-supervised learning methods," *IEEE Transactions on Medical Imaging*, Vol. 39, no. 5, pp. 1438–1447, 2020.

35. Eigen, D., C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Advances in Neural Information Processing Systems 27* (Ghahramani, Z., M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, eds.), pp. 2366–2374, Curran Associates, Inc., 2014.

36. Liu, F., C. Shen, G. Lin, and I. Reid, "Learning depth from single monocular images using deep convolutional neural fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 38, no. 10, pp. 2024–2039, 2016.

37. Laina, I., C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," *CoRR*, Vol. abs/1606.00373, 2016.

38. Turan, M., *et al.*, "Unsupervised odometry and depth learning for endoscopic capsule robots," *arXiv preprint arXiv:1803.01047*, 2018.

39. Lu, Y., and G. Lu, "Deep unsupervised learning for simultaneous visual odometry and depth estimation," in *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 2571–2575, 2019.

40. Mahmood, F., R. Chen, and N. J. Durr, "Unsupervised reverse domain adaptation for synthetic medical images via adversarial training," *IEEE Transactions on Medical Imaging*, Vol. 37, no. 12, pp. 2572–2581, 2018.

41. Mahmood, F., and N. J. Durr, "Deep learning and conditional random fields-based depth estimation and topographical reconstruction from conventional endoscopy," *CoRR*, Vol. abs/1710.11216, 2017.

42. Garg, R., V. K. B. G, and I. D. Reid, "Unsupervised CNN for single view depth estimation: Geometry to the rescue," *CoRR*, Vol. abs/1603.04992, 2016.

43. Zhang, Y., S. Xu, B. Wu, J. Shi, W. Meng, and X. Zhang, "Unsupervised multi-view constrained convolutional network for accurate depth estimation," *IEEE Transactions on Image Processing*, Vol. 29, pp. 7019–7031, 2020.

44. Yin, Z., and J. Shi, "Geonet: Unsupervised learning of dense depth, optical flow and camera pose," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1983–1992, 2018.

45. Chen, R., T. Bobrow, T. Athey, F. Mahmood, and N. Durr, "Slam endoscopy enhanced by adversarial depth prediction," 06 2019.

46. Jiao, J., J. Jiao, Y. Mo, W. Liu, and Z. Deng, "Magicvo: End-to-end monocular visual odometry through deep bi-directional recurrent convolutional neural network," *CoRR*, Vol. abs/1811.10964, 2018.

47. Meng, X., C. Fan, and Y. Ming, "Visual odometry based on convolutional neural networks for large-scale scenes." EasyChair Preprint no. 413, EasyChair, 2018.

48. Mountney, P., D. Stoyanov, A. Davison, and G.-Z. Yang, "Simultaneous stereoscope localization and soft-tissue mapping for minimal invasive surgery," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 347–354, Springer, 2006.

49. Stoyanov, D., M. V. Scarzanella, P. Pratt, and G.-Z. Yang, "Real-time stereo reconstruction in robotically assisted minimally invasive surgery," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 275–282, Springer, 2010.

50. Lin, B., A. Johnson, X. Qian, J. Sanchez, and Y. Sun, "Simultaneous tracking, 3d reconstruction and deforming point detection for stereoscope guided surgery," in *Augmented Reality Environments for Medical Imaging and Computer-Assisted Interventions*, pp. 35–44, Springer, 2013.

51. Mirota, D. J., H. Wang, R. H. Taylor, M. Ishii, G. L. Gallia, and G. D. Hager, "A system for video-based navigation for endoscopic endonasal skull base surgery," *IEEE transactions on medical imaging*, Vol. 31, no. 4, pp. 963–976, 2011.

52. Chen, R. J., T. L. Bobrow, T. L. Athey, F. Mahmood, and N. J. Durr, "Slam endoscopy enhanced by adversarial depth prediction," *ArXiv*, Vol. abs/1907.00283, 2019.

53. Turan, M., Y. Almalioglu, H. Araujo, E. Konukoglu, and M. Sitti, "A non-rigid map fusion-based direct slam method for endoscopic capsule robots," *International journal of intelligent robotics and applications*, Vol. 1, no. 4, pp. 399–409, 2017.

54. Turan, M., Y. Pilavci, R. Jamiruddin, H. Araujo, E. Konukoglu, and M. Sitti, "A fully dense and globally consistent 3d map reconstruction approach for gi tract to enhance therapeutic relevance of the endoscopic capsule robot," 05 2017.

55. Hong, S. P., J. Cheon, T. Kim, S. Song, and W. Kim, "Comparison of the diagnostic yield of "mirocam" and "pillcam sb" capsule endoscopy," *Hepato-gastroenterology*, Vol. 59, pp. 778–81, 05 2012.

56. Artec3D, "User handbook - artec3d eva." Last accessed August 2019.

57. 3D, S., "User handbook - shining 3d einscan." Last accessed August 2019.

58. Incetan, K., I. O. Celik, A. Obeid, G. I. Gokceler, K. B. Ozyoruk, Y. Almalioglu, R. J. Chen, F. Mahmood, H. Gilbert, N. J. Durr, and M. Turan, "Vr-caps: A virtual environment for capsule endoscopy," 2020.

59. Tsai, R. Y., and R. K. Lenz, "A new technique for fully autonomous and efficient 3d robotics hand/eye calibration," *IEEE Transactions on Robotics and Automation*, Vol. 5, pp. 345–358, June 1989.

60. Ou-Yang, M., Y.-L. Chen, H.-H. Lee, S.-c. LU, and H.-M. Wu, "Wide-angle lens for miniature capsule endoscope," pp. 608010–608010, 02 2006.

61. Chen, H. S., and Y.-H. Lin, "An endoscopic system adopting a liquid crystal lens with an electrically tunable depth-of-field," *Optics express*, Vol. 21, pp. 18079–88, 07 2013.

62. Pertuz, S., "Defocus simulation," *MATLAB Central File Exchange.*, June 7, 2020.

63. Bian, J.-W., *et al.*, "Unsupervised scale-consistent depth and ego-motion learning from monocular video," 2019.

64. Ranjan, A., V. Jampani, K. Kim, D. Sun, J. Wulff, and M. J. Black, "Adversarial collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation," *CoRR*, Vol. abs/1805.09806, 2018.

65. Horn, B. K., "Closed-form solution of absolute orientation using unit quaternions," *Josa a*, Vol. 4, no. 4, pp. 629–642, 1987.

66. Handa, A., T. Whelan, J. McDonald, and A. J. Davison, "A benchmark for rgb-d visual odometry, 3d reconstruction and slam," in *Robotics and automation (ICRA), 2014 IEEE international conference on*, pp. 1524–1531, IEEE, 2014.

67. Pogorelov, K., K. R. Randel, T. de Lange, S. L. Eskeland, C. Griwodz, D. Johansen, C. Spampinato, M. Taschwer, M. Lux, P. T. Schmidt, M. Riegler, and P. Halvorsen, "Nerthus: A bowel preparation quality video dataset," in *Proceedings of the 8th ACM on Multimedia Systems Conference*, MMSys'17, (New York, NY, USA), pp. 170–174, ACM, 2017.