VIRTUAL ENDOSCOPY TO GENERATE BENCHMARK DATASETS TO AUGMENT ENDOSCOPIC VIDEO ANALYSIS

SARITA SINGH

DEPARTMENT OF ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY DELHI

JULY 2025

VIRTUAL ENDOSCOPY TO GENERATE BENCHMARK DATASETS TO AUGMENT ENDOSCOPIC VIDEO ANALYSIS

by

SARITA SINGH

Department of Electrical Engineering

Submitted

in partial fulfillment of the requirements of the degree of

Doctor of Philosophy

to the

INDIAN INSTITUTE OF TECHNOLOGY DELHI

JULY 2025

Certificate

This is to certify that the thesis entitled **"Virtual endoscopy to generate benchmark datasets to augment endoscopic video analysis"**, submitted by **Sarita Singh** to the Department of Electrical Engineering, Indian Institute of Technology Delhi, for the award of the degree of **Doctor of Philosophy** in July 2025, is a record of the original, bona fide research work carried out by her under our supervision and guidance.

The matter embodied in this thesis has not been submitted in part or in full to any other University or Institute for the award of any degree or diploma.

Prof. Basabi Bhaumik
Professor,
Department of Electrical Engineering,
Indian Institute of Technology Delhi,
Hauz Khas, New Delhi - 110016,
INDIA.

Prof. Shouribrata Chatterjee
Professor,
Department of Electrical Engineering,
Indian Institute of Technology Delhi,
Hauz Khas, New Delhi - 110016,
INDIA.

Acknowledgement

First and foremost, I would like to express my gratitude to my supervisor, Prof. Shouribrata Chatterjee who believed in me and gave me a chance to pursue this degree. I am profoundly grateful to him for his invaluable support, guidance, and all the required resources throughout the research journey.

I would express my sincere gratitude to my supervisor, Prof. Basabi Bhaumik for her indispensable guidance, support, and inspiration throughout these years. She has consistently provided unwavering mentorship, offering solutions for both professional and personal challenges. I cannot count the number of times Prof. Bhaumik has offered insightful discussions and advices, helping me navigate through the challenges encountered in my research. I feel truly fortunate to have had the opportunity to work under her guidance over these years. My work and accomplishments were possible because of her consistent support and encouragement. I would also like to express my gratitude to Prof. Pankaj Sharan for the expert advice on the mathematical formulations of a section presented in this work.

I express my deep gratitude to mu SRC members Prof. Mukul Sarkar, Prof, Swades De and Prof. Ankur Gupta, for their insightful feedback and invaluable guidance during my SRC presentation. I am profoundly grateful to Prof. Jayadeva for providing the necessary resources, the servers equipped with GPUs, for successfully executing the machine learning processes.

I express my gratitude to Dr. Deepak Gunjan of All India Institute of Medical Sciences India for providing wireless capsule endoscope records data from his lab for the study. I would also like to acknowledge the contributions of my colleagues from the Indian Institute of Technology Delhi, Sarthak Gupta and Vaibhav Soni, for their support during the initial stages of this research project.

I am immensely grateful to my family members, particularly my father and mother, and my husband, for their blessings, and unwavering assistance during the course of my research. I am deeply indebted for their moral support and handling of numerous social responsibilities, which helped me remain focused on my work. This research contribution would not have been possible without their assistance, and words alone are not sufficient to express my gratitude.

(Sarita Singh)

Abstract

The current state of medical technology has demonstrated a convergence towards minimally invasive diagnosis and treatment. In an effort to mitigate the patient's distress during the screening of gastrointestinal tract using traditional wired endoscopy, researchers have introduced the wireless capsule endoscopes (WCEs). The capsule is a pill-shaped swallowable device equipped with a camera, the patient ingests. It travels through the gastrointestinal tract by virtue of the natural mechanism of peristalsis and captures images of the inner lining. An accurate localization method is required for the decision of drug administration and other therapeutic procedures as it depends on the location of the ailment.

The image-based localization techniques estimate the movement of the wireless capsule endoscope by processing the image sequence captured by the capsule camera without utilizing any additional device. Deep learning based techniques have an immense potential to improve endoscopic video analysis, specifically the pose estimation of endoscope camera and simultaneous localization and mapping. At present, the absence of large, varied, and accurately annotated datasets of the gastrointestinal tract limits the efficient learning of deep learning-based models and their effective performance evaluation.

To address the lack of accurately annotated datasets of the gastrointestinal tract, we have developed SimIntestine, the virtual models of two major sections of the gastrointestinal tract: small intestine and large intestine. The virtual models closely mimic the anatomical structure and motility of functioning intestines inside a living human being. The models incorporate the distinctive characteristics of these organs: the intricate shape with various degrees of bends, plicae circulares, villi, deformation of the walls due to peristalsis, brush border-like appearance of the small intestine, haustral folds in the large intestine, their unique surface textures, etc. We have also developed virtual capsule endoscopes for both small and large intestines. These endoscopes navigate through the virtual models of the intestines and provide color images and depth maps. The purpose of implementing various physiological properties of the organs in detail is to produce virtual capsule endoscope images that closely resemble and are almost indistinguishable from real endoscope images. This allows the synthetic dataset to be used for transfer learning in supervised as well as unsupervised learning-based depth and ego-motion detection algorithms for real endoscope image datasets.

Using SimIntestine, we trained the state-of-the-art unsupervised learning based network architectures, Endo-SfMLearner[1] and Monodepth2 [2] for depth estimation using the synthetic image dataset; and assessed its performance on images from KID [3] dataset and REAL-Colon [4] dataset. The results show that the models trained on our synthetic datasets works remarkably well on real endoscope images. We also used the techniques Endo-SfMLearner [1] and Monodepth2 [2] for the evaluation of pose estimation using the synthetic images from the virtual small and large

intestine models. At the end, we used Endo-Depth-and-Motion [5] for the 3D reconstruction of the small intestine from the synthetic images. The results demonstrate the utilization of synthetic datasets for evaluating depth and pose estimation, and 3D reconstruction techniques.

In addition, we modeled some of the health conditions, polyp, bleeding, and ulcerative colitis, for rendering images of affected regions of the intestines to aid disease detection and classification algorithms. The proposed framework provides a comprehensive and physically realistic annotated synthetic dataset benchmark of intestines which can be used to improve endoscopic video analysis, specifically in the domain of pose estimation and simultaneous localization and mapping which is challenging to obtain using real endoscope unannotated dataset.

This research is intended to help physicians locate the wireless capsule endoscope's location inside the gastrointestinal tract and find the location of the ailment detected in the endoscopic images. Our datasets with ground truth position in virtual small intestines is expected to promote the development of real-time feedback systems to locate the capsule endoscope's position inside the gastrointestinal tract and determine the location of the abnormalities detected. The comprehensive and physically realistic annotated synthetic dataset captures the deformation of the wall in the gastrointestinal tract due to peristalsis and provides a benchmark which can be used to improve the ego-motion estimation techniques for endoscopic videos.

सार

चिकित्सा प्रौद्योगिकी की वर्तमान स्थिति ने न्यूनतम आक्रामक निदान और उपचार की ओर अभिसरण प्रदर्शित किया है। पारंपरिक वायर्ड एंडोस्कोपी द्वारा जठरांत्र संबंधी मार्ग की जाँच के दौरान रोगी की परेशानी को कम करने के प्रयास में, शोधकर्ताओं ने वायरलेस कैप्सूल एंडोस्कोप (WCE) विकसित किए हैं। यह कैप्सूल एक गोली के आकार का निगलने योग्य उपकरण है जिसमें एक कैमरा लगा होता है। यह क्रमाकुंचन की प्राकृतिक क्रियाविधि द्वारा जठरांत्र संबंधी मार्ग से होकर गुजरता है और आंतरिक परत के चित्र लेता है। दवा के प्रशासन और अन्य चिकित्सीय प्रक्रियाओं के निर्णय के लिए एक सटीक स्थानीयकरण विधि की आवश्यकता होती है क्योंकि यह रोग के स्थान पर निर्भर करता है।

छवि-आधारित स्थानीयकरण तकनीकें, बिना किसी अतिरिक्त उपकरण का उपयोग किए, कैप्सूल कैमरे द्वारा कैप्चर किए गए छवि अनुक्रम को संसाधित करके वायरलेस कैप्सूल एंडोस्कोप की गति का अनुमान लगाती हैं। डीप लर्निंग-आधारित तकनीकों में एंडोस्कोपिक वीडियो विश्लेषण, विशेष रूप से एंडोस्कोप कैमरे की मुद्रा का अनुमान और समकालिक स्थानीयकरण एवं मानचित्रण, को बेहतर बनाने की अपार क्षमता है। वर्तमान में, जठरांत्र संबंधी मार्ग के बड़े, विविध और सटीक रूप से एनोटेट किए गए डेटासेट का अभाव डीप लर्निंग-आधारित मॉडलों के कुशल शिक्षण और उनके प्रभावी प्रदर्शन मूल्यांकन को सीमित करता है।

जठरांत्र संबंधी मार्ग के सटीक रूप से एनोटेट किए गए डेटासेट की कमी को दूर करने के लिए, हमने सिमइंटेस्टाइन विकसित किया है, जो जठरांत्र संबंधी मार्ग के दो प्रमुख भागों: छोटी आंत और बड़ी आंत के आभासी मॉडल हैं। ये आभासी मॉडल जीवित मानव के अंदर कार्यशील आंतों की शारीरिक संरचना और गतिशीलता की हूबहू नकल करते हैं। ये मॉडल इन अंगों की विशिष्ट विशेषताओं को समाहित करते हैं: विभिन्न स्तरों के मोड़ों वाला जटिल आकार, गोलाकार प्लिके, विली, क्रमाकुंचन के कारण दीवारों का विरूपण, छोटी आंत का ब्रश बॉर्डर जैसा रूप, बड़ी आंत में हस्ट्रल सलवटें, उनकी अनूठी सतही बनावट, आदि। हमने छोटी और बड़ी दोनों आंतों के लिए आभासी कैप्सूल एंडोस्कोप भी विकसित किए हैं। ये एंडोस्कोप आंतों के आभासी मॉडलों में नेविगेट करते हैं और रंगीन चित्र और गहराई मानचित्र प्रदान करते हैं। अंगों के विभिन्न शारीरिक गुणों को विस्तार से लागू करने का उद्देश्य आभासी कैप्सूल एंडोस्कोप चित्र तैयार करना है जो वास्तविक एंडोस्कोप चित्रों से लगभग मिलते-जुलते और लगभग अप्रभेद्य हों। इससे सिंथेटिक डेटासेट को वास्तविक एंडोस्कोप छवि डेटासेट के लिए पर्यवेक्षित और साथ ही अपर्यवेक्षित शिक्षण-आधारित गहराई और अहं-गति पहचान एल्गोरिदम में स्थानांतरण शिक्षण के लिए उपयोग करने की अनुमति मिलती है।

सिमइंटेस्टाइन का उपयोग करते हुए, हमने सिंथेटिक इमेज डेटासेट का उपयोग करके गहराई का अनुमान लगाने के लिए अत्याधुनिक अनसुपरवाइज्ड लर्निंग-आधारित नेटवर्क आर्किटेक्चर, एंडो-एसएफएमलर्नर[1] और

मोनोडेप्थ2 [2] को प्रशिक्षित किया; और केआईडी [3] डेटासेट और रियल-कॉलन [4] डेटासेट की छवियों पर इसके प्रदर्शन का आकलन किया। परिणाम दर्शाते हैं कि हमारे सिंथेटिक डेटासेट पर प्रशिक्षित मॉडल वास्तविक एंडोस्कोप छवियों पर उल्लेखनीय रूप से अच्छा काम करते हैं। हमने आभासी छोटी और बड़ी आंत के मॉडल से सिंथेटिक छवियों का उपयोग करके पोज़ अनुमान के मूल्यांकन के लिए एंडो-एसएफएमलर्नर [1] और मोनोडेप्थ2 [2] तकनीकों का भी उपयोग किया। अंत में, हमने सिंथेटिक छवियों से छोटी आंत के 3डी पुनर्निर्माण के लिए एंडो-डेप्थ-एंड-मोशन [5] का उपयोग किया।

इसके अलावा, हमने रोग पहचान और वर्गीकरण एल्गोरिदम में सहायता के लिए आंतों के प्रभावित क्षेत्रों की छवियों को प्रस्तुत करने हेतु कुछ स्वास्थ्य स्थितियों, जैसे पॉलीप, रक्तस्राव और अल्सरेटिव कोलाइटिस का मॉडल तैयार किया। प्रस्तावित ढाँचा आंतों का एक व्यापक और भौतिक रूप से यथार्थवादी एनोटेटेड सिंथेटिक डेटासेट बेंचमार्क प्रदान करता है जिसका उपयोग एंडोस्कोपिक वीडियो विश्लेषण को बेहतर बनाने के लिए किया जा सकता है, विशेष रूप से पोज़ अनुमान और समकालिक स्थानीयकरण और मानचित्रण के क्षेत्र में, जिसे वास्तविक एंडोस्कोप अननोटेटेड डेटासेट का उपयोग करके प्राप्त करना चुनौतीपूर्ण है।

इस शोध का उद्देश्य चिकित्सकों को जठरांत्र संबंधी मार्ग के अंदर वायरलेस कैप्सूल एंडोस्कोप की स्थिति का पता लगाने और एंडोस्कोपिक छवियों में पाई गई बीमारी का स्थान निर्धारित करने में मदद करना है। आभासी छोटी आंत में वास्तविक स्थिति वाले हमारे डेटासेट से जठरांत्र संबंधी मार्ग के अंदर कैप्सूल एंडोस्कोप की स्थिति का पता लगाने और पाई गई असामान्यताओं का स्थान निर्धारित करने के लिए रीयल-टाइम फीडबैक सिस्टम के विकास को बढ़ावा मिलने की उम्मीद है। व्यापक और भौतिक रूप से यथार्थवादी एनोटेटेड सिंथेटिक डेटासेट, क्रमाकुंचन के कारण जठरांत्र संबंधी मार्ग की दीवार के विरूपण को दर्शाता है और एक बेंचमार्क प्रदान करता है जिसका उपयोग एंडोस्कोपिक वीडियो के लिए अहंकार-गति अनुमान तकनीकों को बेहतर बनाने के लिए किया जा सकता है।

Contents

List of Figures

List of Tables

Abbreviations

GI	**G**astro **I**ntestinal
CRC	**C**olo **R**ectal **C**ancer
WCE	**Wireless Capsule Endoscopy**
CMOS	**C**omplementary **M**etal **O**xide **S**emiconductor
LED	**L**ight **E**mitting **D**iode
UHF	**U**ltra **H**igh **F**requency
FDA	**F**ood and **D**rug **A**dministration
CCD	**C**harge **C**oupled **D**evice
RF	**R**adio **F**requency
SLAM	**S**imulatneous **L**ocalization **A**nd **M**apping
FCC	**F**ederal **C**ommunication **C**ommission
MICS	**M**edical **I**mplant **C**ommunication **S**ervices
DoA	**D**irection **o**f **A**rrival
ToA	**T**ime **o**f **A**rrival
TDoA	**T**ime **D**ifference **o**f **A**rrival
RFID	**R**adio **F**requency **ID**entification
RSSI	**R**eceived **S**ignal **S**trength **I**ndicator
IMU	**I**nertial **M**easurement **U**nit
HFSS	**H**igh **F**requency **S**tructure **S**imulator
SAR	**S**pecific **A**bsorption **R**ate
SIFT	**S**cale-**I**nvariant **F**eature **T**ransform
ASIFT	**A**ffine **S**cale-**I**nvariant **F**eature **T**ransform
SURF	**S**peeded **U**p **R**obust **F**eatures
FLANN	**F**ast **L**ibrary for **A**pproximate **N**earest **N**eighbors
ORB	**O**riented FAST and **R**otated **B**RIEF
MSAC	**M**-estimator **SA**mple **C**onsensus
RANSAC	**Ran**dom **SA**mple **C**onsensus
KLT	**K**anade-**L**ucas-**T**omasi

SfS	**S**hape **f**rom **S**hading
SfM	**S**tructure **f**rom **M**otion
MVS	**M**ulti **V**iew **S**tereo
ICP	**I**terative **C**losest **P**oint
RAPID	**R**eporting **A**nd **P**rocessing of **I**mages and **D**ata
CNN	**C**onvolutional **N**eural **N**etwork
RNN	**R**ecurrent **N**eural **N**etwork
SB	**S**mall **B**owel
MDR	**M**odel of **D**eformable **R**ings
PVC	**P**oly**v**inyl **C**hloride
CT	**C**omputed **T**omography
DICOM	**D**igital **I**maging and **Com**munications in **M**edicine
VR	**V**irtual **R**eality
SimInt	**Sim**ulated **Int**estine
MMC	**M**igrating **M**otor **C**omplex
PC	**P**licae **C**irculares
STL	**St**ereo**l**ithography
DispNet	**Disp**arity Estimation **Net**work
PoseNet	**Pose** Estimation **Net**work
BSDF	**B**idirectional **S**cattering **D**istribution **F**unction
LIDAR	**LI**ght **D**etection **A**nd **R**anging
RMSE	**R**oot **M**ean **S**quare **E**rror
Abs Rel	**Abs**olute **Rel**ative
MAE	**M**ean **A**bsolute **E**rror
log RMSE	**log**arithmic **R**oot **M**ean **S**quare **E**rror
GIANA	**G**astrointestinal **I**mage **ANA**lysis
DOF	**D**egrees **O**f **F**reedom
ATE	**A**bsolute **T**rajectory **E**rror
trans RPE	**trans**lational **R**elative **P**ose **E**rror
rot RPE	**rot**ational **R**elative **P**ose **E**rror
SSIM	**S**tructural **S**imilarity **I**ndex **M**easure
PSSIM	**P**atch **S**tructural **S**imilarity **I**ndex **M**easure
SC	**S**cale **C**onsistency
ESAB	**E**ndoSfMLearner **S**patial **A**ttention **B**lock
SO	Special Orthogonal Group
SE	Special Euclidean Group
TSDF	**T**runcated **S**igned **D**istance **F**unction

Symbols

a	radius of the tube
A, B	points along the circumference of plicae circularis
$A_{mod}(t)$	amplitude of segmentation wave function $r_{seg}(d,t)$
$A_p(t)$	amplitude of the gaussian functions at time t
$A_{peris}(t)$	amplitude of the peristalsis wave
A_{seg}	amplitude of segmentation wave
B	magnetic flux intensity
c	velocity of wave
$d(n,k)$	distance from pylorus to k^{th} point of the n^{th} segment
$d_s(x,y,z)$	signed distance from voxel at (x, y,z) to the nearest surface
$D_t(p)$	predicted depth at pixel p
E	essential matrix
$\hat{E}_s$	explainability prediction network
$f_r(d,t)$	sinusoidal waveforms at distance d at time t
F	fundamental matrix
$g_r(d,t)$	sinusoidal waveforms (half wavelength apart) at distance d at time t
$\tilde{h}$	radial displacement of walls from center line
I_s	source image
I_t	target image
$\hat{I}_s$	synthesized image generated after warping source image I_s
$J_n(x)$	zeros of bessel function
$kx(i), ky(i), kz(i)$	coordinates of beginning node of i^{th} segment
K	camera intrinsics matrix
l_{AB}	length of section of plicae circularis between points A and B
l_s	distance between two adjacent points on the medial axis
L	path loss
L_{geo}	geometry consistency loss
$L_{\text{min-reproj}}$	minimum reprojection loss

$L_{\text{multi-scale}}$	multi-scale loss
L_{photo}	photometric loss
L_{reg}	regularization term
L_{smooth}	smoothness loss
L_{vs}	view synthesis loss
$\vec{m}$	magnetic dipole moment vector
n	total number of frames
nk	order of bessel function
n_s	number of segments in model
$n_{p(i)}$	number of points in i^{th} segment
N	number of folds in plicae circularis
p	pixel
p'	pixel after reconstruction
$p(t)$	controls the checkered appearance of segmentation wave
pk	number of zeros of bessel function
$\mathrm{P}_i^{\text{est}}$	estimated position at i^{th} frame
P_i^{gt}	ground truth position at i^{th} frame
$q(d)$	controls the placement of plicae circulares
r_i	radius of inner tubular structure of intestine
R_i	radius of outer tubular structure of intestine
r_i'	radius of inner tubular structure of intestine after contraction
R_i'	radius of outer tubular structure of intestine after contraction
r_1	radius of outer two quarter loops of plicae circularis after contraction
r_2	radius of inner semicircle loop of plicae circularis after contraction
$\vec{r}$	position vector
$r_{pc}(d)$	radial displacement of intestine due to plicae circulares
$r_{peris}(d,t)$	radial displacement of intestine due to peristalsis wave
$r_{seg}(d,t)$	radial displacement of intestine due to segmentation wave
$r_{si}(d)$	radial displacement of intestine due to peristalsis and segmentation
R	rotation matrix
$\mathrm{R}_i^{\text{est}}$	estimated rotation matrix at i^{th} frame
R_i^{gt}	ground truth rotation matrix at i^{th} frame
s	scale of source image
s_i	spacing between two plicae circulares of i^{th} segment
S	shadowing factor
t	time
t_{pc}	thickness of plicae circularis
$\tilde{t}$	time

$\mathrm{t}_i^{\mathrm{est}}$	estimated translation vector at i^{th} frame
$\mathrm{t}_i^{\mathrm{gt}}$	ground truth translation vector at i^{th} frame
T	translation matrix
T_s	time period of the segmentation wave
$T_{t\rightarrow t+1}$	6-DOF relative pose matrix from image t to $t+1$
T_β	affine brightness transformation
v_p	velocity of peristalsis wave
$x(i,j), y(i,j), z(i,j)$	coordinates of j^{th} point of i^{th} segment
$x_{arc}(i), y_{arc}(i), z_{arc}(i)$	magnitudes of arc along x, y, z axes of i^{th} segment
$\tilde{x}$	axial coordinate
X	total section of plicae circularis where there is no fold
w_i	width of plica circularis of i^{th} segment

Greek Symbols

α_t	contrast of image
β_t	brightness of image
γ	path loss component
δ	truncation threshold
λ	wavelength of wave
λ_e	weight for explainability mask regularization
λ_p	spatial width of peristalsis wave
λ_s	wavelength of segmentation wave
λ_{smooth}	weight for depth smoothness loss
μ_0	magnetic permeability of air
∇_x	spatial gradients along x direction
∇_y	spatial gradients along y direction
$\Omega_{pc}(d)$	Gaussian curve for modeling plicae circulares
π	projection function
ϕ	amplitude of the wave
$\phi_1(d,t)$,$\phi_2(d,t)$	spatially displaced Gaussian functions
σ_p	standard deviation of gaussian functions, $\phi_1(d,t)$ and $\phi_2(d,t)$
σ_{pc}	standard deviation of gaussian function, $\Omega_{pc}(d)$
σ_s	standard deviataion of shadowing factor
θ_{s1}, θ_{s2}	points along the circumference of plicae circularis