Anomaly Detection in Endoscopic Videos: Keyframe Extraction to Designing Clinical and Synthetic Datasets

Thesis submitted in partial fulfilment of the requirements
for the award of the degree of

Doctor of Philosophy

in

Computer Science and Engineering

by

Vanshali Sharma

Under the supervision of

Prof. Pradip K. Das and Prof. M.K. Bhuyan

Department of Computer Science and Engineering

Indian Institute of Technology Guwahati

Guwahati - 781039 Assam India

August, 2024

Dedicated to my parents for empowering me to chase my aspirations, my sister for her love and unwavering support and my husband for being a steady beacon of inspiration

Acknowledgements

I express my earnest gratitude to *Prof. Pradip K. Das* and *Prof. M.K. Bhuyan*, my Ph.D. advisors. They have always provided me with a flexible atmosphere and freedom to think independently about research problems, which helped me nurture my research interests. At the same time, their guidance helped me to refrain from unwise directions and made me aware of the right path to reach the ultimate goal. They unfailingly exhibited patience throughout the whole course, from supporting my ideas to the outcomes obtained, no matter success or failure. I am also thankful to my doctoral committee members, *Prof. Shivashankar B. Nair, Dr. Pinaki Mitra*, and *Dr. Manoj Majhi*, for providing valuable feedback during the research.

I would like to thank my collaborators, *Dr. Debesh Jha, Dr. Ulas Bagci* from *Northwestern University, Chicago, USA*, for providing me with the opportunity to benefit from their expertise in the related domain. I am grateful for their suggestions and constructive feedback, which have immensely enriched my knowledge. I would also like to thank our medical collaborators, *Dr. Thomas de Lange* and *Dr. Neethi Dasu*, for their valuable time and efforts in annotating the dataset.

I am also grateful to our collaborator, *Prof. Yuji Iwahori* from *Chubu University, Japan*, for his discussions and to clinical expert *Dr. Kunio Kasugai* from *Aichi Medical University, Japan*, for supporting our work. I would also like to thank my co-author cum motivator, *Dr. Kangkana Bora*, who consistently encouraged me to advance in my career. I am also thankful to my senior and co-author, *Dr. Pradipta Sasmal*, for his valuable input and guidance.

I thank organizations such as *AAAI, Aichi Medical University, Japan* and the *Department of CSE, IITG*, for providing travel grants/funding for conferences and journals. Also, I thank *DST, Government of India*, for providing financial assistantship under the *INSPIRE* scheme. I am also thankful to the *Department of CSE* for providing the necessary computing resources used during the work. I also thank *Bhriguraj Borah* and *Nanu Alan Kachari* for always being approachable for any technical issues.

I want to acknowledge the friendly and supportive research environment created by my senior, *Dr. Bhagath* and my other labmates, *Susma, Komal, Ashita* and *Himanshu*. I am

also fortunate to have several wonderful friends who were there to support and celebrate with me during various events. I especially want to thank *Surabhi, Suraj, Karnish, Dr. Sonia, Dr. Divya, Nidhi, Tathagata, Megha, Shubham, Darshika, Sandeep, Farhan, Dr. Aakansha, Dr. Pallabi* and *Dr. Deepankar*.

I ascribe this accomplishment to my mother, *Mrs. Seema Sharma*, for her constant motivation and for helping my dreams take flight and to my father, *Mr. Rishi Kumar*, for always believing in me and my capabilities. I thank my sister, *Dr. Parnika*, for her immense love and care. Her support has always been a cornerstone in my endeavors. I would also like to extend my heartfelt appreciation and gratitude to my husband, *Dr. Abhishek* for his continuous support and motivation. I would also like to express my gratitude to my in-laws for understanding the importance of this journey.

Declaration

I certify that

- The work contained in this thesis is original and has been done by myself and under the general supervision of my supervisor(s).
- The work reported herein has not been submitted to any other Institute for any degree or diploma.
- Whenever I have used materials (concepts, ideas, text, expressions, data, graphs, diagrams, theoretical analysis, results, etc.) from other sources, I have given due credit by citing them in the text of the thesis and giving their details in the references. Elaborate sentences used verbatim from published work have been clearly identified and quoted.
- I also affirm that no part of this thesis can be considered plagiarism to the best of my knowledge and understanding and take complete responsibility if any complaint arises.
- I am fully aware that my thesis supervisor(s) are not in a position to check for any possible instance of plagiarism within this submitted work.

August 27, 2024 Vanshali Sharma

Department of Computer Science and Engineering
Indian Institute of Technology Guwahati
Guwahati - 781039 Assam India

Prof. Pradip K. Das
Professor
Email : pkdas@iitg.ac.in
Phone : +91-361-258-2523

Prof. M.K. Bhuyan
Professor
Email: mkb@iitg.ac.in
Phone: +91-361-258-2353

Certificate

This is to certify that this thesis entitled "**Anomaly Detection in Endoscopic Videos: Keyframe Extraction to Designing Clinical and Synthetic Datasets**" submitted by **Vanshali Sharma**, in partial fulfilment of the requirements for the award of the degree of Doctor of Philosophy, to the Indian Institute of Technology Guwahati, Assam, India, is a record of the bonafide research work carried out by her under my guidance and supervision at the Department of Computer Science and Engineering, Indian Institute of Technology Guwahati, Assam, India. To the best of my knowledge, no part of the work reported in this thesis has been presented for the award of any degree at any other institution.

Date: August 27, 2024
Place: Guwahati

Prof. Pradip K. Das
(Main-supervisor)

Prof. M.K. Bhuyan
(Co-supervisor)

Abstract

Gastrointestinal (GI) cancers, specifically colorectal cancers (CRC), are prevalent and significant contributors to global cancer-related deaths. CRC originates from pre-malignant polyps, which can be detected through a colonoscopy procedure, during which videos of a patient's colon are captured. However, analyzing screening videos for related diagnosis and treatment faces challenges due to a large proportion of low-quality data, risking human review errors. Further, the low-quality data and the limited availability of large-scale annotated datasets pose significant hurdles in building automated computer-aided diagnostic systems. This thesis addresses these challenges while aligning with standard clinical procedures. To maintain this uniformity, we mimic these manual procedures in our proposed automated pipeline and present solutions to problems encountered at different stages.

A standard clinical analysis of colonoscopy videos generally begins with manually reviewing recordings and gradually confines the analysis to keyframes for retrospective treatments. Hence, initially, this thesis focuses on automating this task to reduce the clinicians' burden. However, the keyframe count could be significantly low in some video recordings captured under extremely unfavorable conditions. Therefore, techniques to extract obscured details of uninformative frames are proposed. Following the pre-processing stage, the thesis addresses issues related to automated diagnostic systems, enhancing lesion detection, localization, segmentation, and classification outcomes. One of the crucial concerns in the literature is the lack of reproducibility and fair comparison across different segmentation techniques due to inconsistent evaluation datasets, as revealed in our case study. Motivated by this, our focus is on resolving dataset availability issues, as a good-quality, diverse dataset enhances lesion detection performance and promotes reproducibility. Thus, this thesis incorporates effective keyframe selection and other pre-processing techniques, meticulous dataset curation, and synthetic image generation. The four significant contributions of the thesis are highlighted below.

First, a multi-stage framework is presented that focuses on *keyframe extraction* to select good-quality, non-redundant frames and enforce diversity in the final frames for analysis. The framework enhances polyp detection and polyp localization outcomes while reducing

processing time. Our novel multi-scale attention-based localization model, *YcOLOn*, further improves the localization task within the framework.

Second, we propose two approaches to overcome the limitations of our keyframe extraction framework. These techniques focus on obtaining obscured clinical details from uninformative frames with artifacts. One method is an *adversarial-based* approach that focuses on translating uninformative frames into clinically significant frames. This helps in improved polyp localization. With a similar aim to deal with artifacts, a *DWT-based encoder-decoder* architecture is designed to segment specular highlights while overcoming the issue of overexposed regions in the colonoscopy images.

Third, a *case study* is presented that analyzes different polyp and instrument segmentation algorithms involved in two competitions conducted in the years 2020 and 2021. Such analysis provides an opportunity to compare different state-of-the-art techniques on the same dataset for transparency and reproducibility. Additionally, we release an open-access multi-class dataset, *GastroVision*, for computer-aided diagnosis of GI cancer. It comprises 8000 images from 27 classes covering pathological and normal findings, anatomical landmarks, and cases from therapeutic interventions.

Fourth, two frameworks are proposed to generate synthetic medical images using diffusion models, overcoming the lengthy procedures to acquire real medical datasets. The first framework, *ControlPolypNet*, leverages easily accessible non-polyp frames and converts them into hard-to-find polyp images. The generated polyp images are utilized to augment a real dataset to perform a downstream task of polyp segmentation. The second framework, *PathoPolyp-Diff*, is a text-controlled model to generate diverse polyp images covering different pathologies, imaging modalities, and quality. The generated images are used to augment real datasets to enhance pathology based polyp classification. Further, cross-class label learning is introduced, which learns features from other classes without additional annotations.

Contents

List of Figures

List of Tables

List of Symbols

A	Text prompt without cross-class label learning. xxiii, xxvii, 146, 147, 149, 151, 153, 154
$\mathfrak{a}$	uninformative colonoscopy frame. 67, 68
$\mathcal{A}$	Domain representing uninformative frames. 67
a_l	Latent representation of an image 'a'. 129
a'_l	Reconstructed output of latent representation a_l. 129
α	weighted score used with confidence score. 50
a_t	Noisy image obtained during forward diffusion process on an image 'a'. 128
B	Text prompt with cross-class label learning. xxiii, xxvii, 146, 147, 149, 151, 153–155
b'	Task-specific conditional text prompt. 129
b	Text prompt condition. 129, 131
bb	Two-dimensional coordinates of bounding box containing polyp. 47, 49
$\mathfrak{b}$	uninformative colonoscopy frame. 67, 68
$\mathcal{B}$	Domain representing informative frames. 67
β	weighted score used with polyp-to-frame ratio. 50
$\mathfrak{b}'$	Number of histogram bins. 84
C	Number of channels of an image or feature map. 49
$\mathfrak{C}$	Clusters of redundant frames. 46, 47

c	Object class. 28, 37, 49
cs	Confidence score related to polyp detection. 49
$\mathcal{C}$	Zero convolution operation. 129
D'	Colonoscopy video/image dataset. 45, 47
D_A	Discriminator aiming to distinguish the translated and generated uninformative images from the real uninformative frames. xix, 67, 68
D_B	Discriminator aiming to distinguish the translated and generated informative images from the real informative frames. xix, 67, 68
D	Decoder. 129
D_P	A subset of colonoscopy video/image test set containing only polyp frames. 45
D'_{test}	Colonoscopy video/image test set. 45
D'_{train}	Colonoscopy video/image train set. 45
D'_{val}	Colonoscopy video/image validation set. 45
E'	Output of siamese network for one image. 46
E	Encoder. 129, 131
f	Frames in a colonoscopy video. 24, 44–46
fft	Mean of magnitude spectrum of FFT. 44, 45
fn	Number of false negative samples. 36
fp	Number of false positive samples. 36, 37
G	Global channel context. 48, 49
G_{AB}	Generator aiming to translate uninformative frames to informative frames. 67, 68
γ	Weighted score used with the parameter that is computed to measure the distance from the center of a bounding box. 50

γ'	A factor used for smoothing purposes in dice loss. 77
G_{BA}	Generator aiming to translate informative frames to uninformative frames. 67, 68
Gr	Region in the ground truth mask. 36
H	Height dimension of an image or feature map. 49
h	Height dimension of bounding box around region of interest. 28, 49, 50, 84
$\mathcal{H}$	ControlNet network blocks. 129
I	Raw RGB image. 29, 31, 72, 77, 84
$\mathfrak{f}$	Filter used during DWT procedure. 74, 75
$\mathfrak{i}_{hh}$	High-high sub-band image obtained using DWT. 74, 75
$\mathfrak{i}_{hl}$	High-low sub-band image obtained using DWT. 74, 75
$\mathfrak{i}_{lh}$	Low-high sub-band image obtained using DWT. 74, 75
$\mathfrak{i}_{ll}$	Low-low sub-band image obtained using DWT. 74, 75
$\hat{\mathfrak{i}}_{ll}$	Modified low-low frequency sub-band obtained by performing element-wise multiplication on each value. 75
J	Eucledian distance. 46
k	Number of keyframes. 24
L	Local channel context. 48, 49
M	Set of polyp images overlapped with ground truth mask. 130, 131

$\mathfrak{m}$	Number of frames/images in a raw colonoscopy video/dataset. 24, 29, 31, 72
N	Set of non-polyp images. 130
N'	Set of non-polyp images overlapped with user-defined mask. 130, 135
O	Ground Truth mask. 29, 31, 72, 77
$\hat{O}_i$	Specularity segmentation output. 72
o_l	Feature map at current layer. 44
P	Set of polyp images. 130
p	Patient/case identification number. 45, 47
P'	Set of synthetic polyp images. 130, 135
ϕ	ControlNet network blocks. 129
p_mod	Imaging modality (NBI/WLI). xxvii, 146, 147
Pr	Region in the predicted mask. 36
p_type	Pathological polyp class (adenomatous/hyperplastic). xxvii, 146, 147, 155
Q	A colonoscopy image set. 29, 31, 72, 77
q	Previous layer's feature map. 44
S	Dissimilarity score obtained using Siamese model. 46
T	Threshold. 46
ϵ_θ	Denoising framework to predict the denoised image version. 128, 129, 131
t	Time step uniformly sampled between [1,T']. 128, 129, 131

T'	Noise steps involved during diffusion process. 128
tn	Number of true negative samples. 36, 37
tp	Number of true positive samples. 36
u	x-coordinate of point representing center of bounding box around region of interest. 28, 49, 50
V	A colonoscopy video. 24, 36, 44, 45
$V_{insignificant}$	Insignificant/Uninformative frames in a colonoscopy video. 44
v	y-coordinate of point representing center of bounding box around region of interest. 28, 49, 50
V_{key}	Set of keyframes. 24
V_{NP_frames}	Non-polyp colonoscopy frames. 45
vol	Focus measure of VoL. 44, 45
V_{P_frames}	Colonoscopy frames with polyp. 45
$V_{Stage\text{-}I}$	Colonoscopy video frame set comprising informative frames. 44, 45
$V_{Stage\text{-}III}$	Colonoscopy video frame set comprising clusters of redundant polyp frames. 47
$V_{Stage\text{-}II}$	Colonoscopy video frame set comprising clinically significant polyp frames. 45, 47
W	Width dimension of an image or feature map. 49
w	Width dimension of bounding box around region of interest. 28, 49, 50
w_l	Convolution weight. 44
x	Original dataset size or number of real images used. 135, 137, 146, 147, 149

χ	Chi-square distance. 84
y	Class label. 45–47
z	Margin value used in siamese network. 46
$\mathcal{Z}$	Text encoder. 129, 131

List of Abbreviations

ADR	Adenoma Detection Rate. 21, 22
AFF	Attention Feature Fusion. xviii, 48, 49, 55
AFI	Autofluorescence Imaging. xvii, 22
AI	Artificial Intelligence. xvii, 2, 12, 87, 89–91, 94, 104, 106, 110, 111, 118, 159
AIANs	American Indians and Alaska Natives. 18
AP	Average Precision. 37
AUPRC	Area under Precision-Recall Curve. xviii, 37, 52, 53, 56
BA	Balanced Accuracy. 13, 37, 125, 140, 141, 146
BCE	Binary Cross Entropy. 46, 77
BLI	Blue Light Imaging. 22, 23
BN	Batch Normalization. 49, 73, 74
CAD	Computer-Aided Diagnosis. xvii, 1–4, 6, 7, 104, 108, 124, 159, 161
CD	distance from image's center. 49, 50, 59
CNN	Convolutional Neural Network. 44–46, 71, 73, 74, 76, 77, 85, 98, 99, 117
CRC	Colorectal Cancer. xvii, 2, 4, 5, 17, 18, 21–24, 40
CS	Confidence Score. xviii, xix, 47, 49–52, 54, 56, 58, 59, 61
CSP	Cold Snare Polypectomy. 23
CT	Computed Tomography. 2, 16, 17

DM Diffusion Models. 128

DWT Discrete Wavelet Transform. xix, 10, 11, 63, 64, 67, 71, 73–78, 85, 160

FFT Fast Fourier Transform. 44, 45, 51

FICE Flexible Spectral Imaging Color Enhancement. 22, 23

FID Frechet Inception Distance. 126, 133, 134

FIT Fecal Immunochemical Test. 17

FOBT Fecal Occult Blood Test. 17

FPS Frames Per Second. 89, 93, 94, 97, 100, 101, 106

GAN Generative Adversarial Network. 12, 34, 35, 66, 68, 85, 98, 99, 106, 107, 125, 160

GI Gastrointestinal. xvii, xx, xxv, 3, 4, 11, 13, 15–17, 32–34, 87–92, 94, 110–112, 114, 115, 117, 118, 120, 121, 159

GUI Graphical User Interface. xix, 10, 40, 43, 60–62, 159

HD-WLI High-definition White Light Imaging. 22

HSP Hot Snare Polypectomy. 23

IoU Intersection over Union. 36

JI Jaccard Index. xxi, 12, 36, 124, 133, 136, 156

KID Kernel Inception Index. xxvi, 141, 142

LCI Linked Color Imaging. 22, 23

LDM Latent Diffusion Models. 128

mAP mean Average Precision. 37, 40, 57, 62

MCC	Matthews Correlation Coefficient. 116
MRI	Magnetic Resonance Imaging. 2
MS-CAM	Multi-scale Channel Attention. xviii, 48, 49
NBI	Narrow Band Imaging. xvii, xxii, xxiii, xxvii, 11, 13, 22, 23, 32, 33, 35, 36, 111, 124, 125, 127, 128, 139, 144–147, 149, 151, 153–156, 161
NICE	NBI International Colorectal Endoscopic. xvii, 19, 20
PANet	Path Aggregation Network. 47
PI	Polyp-to-frame ratio. xviii, xix, 41, 49, 50, 54, 59
PSNR	Peak Signal-to-Noise Ratio. 126, 133, 134
PVT	Pyramid Vision Transformer. 31
ROI	Region of Interest. xix, 2, 40, 47, 49, 50, 58, 60, 96
SD	Stable Diffusion Models. 128–131, 139
SGD	Stochastic Gradient Descent. 50, 78, 98, 99, 106
SPP	Spatial Pyramid Pooling. 47
SSIM	Structural Similarity Index Measure. 126, 133, 134
VoL	Variance of Laplacian. 44, 45, 51
WCE	Wireless Capsule Endoscopy. 24
WLI	White Light Imaging. xvii, xxii, xxiii, xxvii, 11, 13, 22, 32, 33, 35, 36, 109, 111, 124, 125, 127, 128, 139, 144–149, 153–156, 161
WPT	Wavelet Packet Transform. 67, 74

1

Introduction

1.1 Overview

Medical image analysis is a field that has evolved over the intersection of healthcare and technology. It leverages fast-growing technical advancements to provide automated image-based solutions to routine clinical challenges. These challenges include but are not limited to operator dependency in decision-making that ultimately relies on the skill set of the concerned physician, manual measurements that are prone to detection errors, and the unavailability of domain experts. As stated by ACR's Data Science Institute, *"Every patient has annual follow-up exams, which can add up to about 100 exams per day for a musculoskeletal radiologist who works with arthroplasty surgeons."*[1] Manual examination and reviewing of such enormous image data involves high risk of errors, delays in treatment, and increase in lesion[2] miss-rate. Thus, it becomes crucial to adopt automated tools to mitigate the above-mentioned challenges.

The concept of automation in medical image analysis introduced Computer-Aided Diagnosis (CAD) systems [1]. These systems are built on computational algorithms which enable the automatic interpretation of medical images for accurate predictions. CAD is further endowed with rapidly evolving machine learning and deep learning techniques. Consequently, in a short span of a few years, automation has revamped the medical imaging domain, resulting in improved disease diagnosis, treatment planning, and patient outcomes. The related research progress and reshaping of technologies can be inferred from Fig. 1.1. This shows that integrating CAD with medical image analysis can provide decision sup-

[1] https://healthitanalytics.com/news/top-5-use-cases-for-artificial-intelligence-in-medical-imaging
[2] a region in an organ or tissue which has suffered damage through injury or disease.

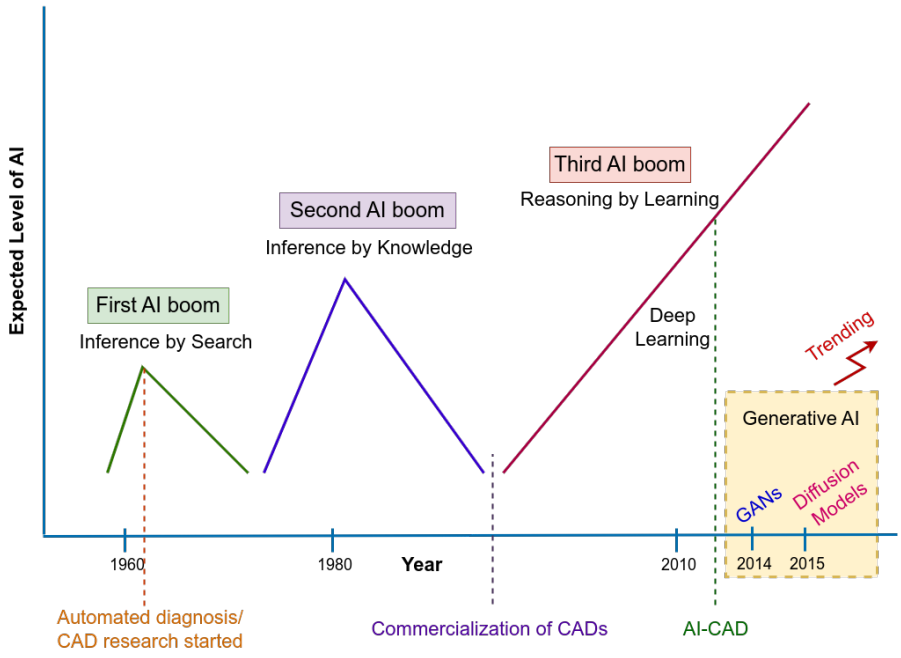


Figure 1.1: CAD research progress and its relation with AI booms and trending technologies.

port in various clinical activities, thus functioning as the second eye for clinicians. A block diagram for a general CAD system is shown in Fig. 1.2.

In a clinical setting, there could be a wide range of tasks and associated imaging modalities. An imaging modality can be preferred over the other, considering the type of medical condition and the severity stage. For example, Magnetic Resonance Imaging (MRI) is used for the detection of brain tumours, Computed Tomography (CT) scans are preferred for pulmonary disorders and dental conditions, X-rays are suggested for bone fractures and arthritis conditions, endoscopy is done for gastric ulcers and esophagitis, and colonoscopy is performed to detect colon polyps and Colorectal Cancer (CRC). These imaging modalities provide an internal view and structures of the infected region [2]. The CAD systems then process the obtained images to recognize lesion patterns and classify them based on some class (abnormal or normal) distinguishing features.

During the diagnosis process, CAD internally performs various subtasks that include pre-processing of the acquired images, detection/segmentation [3] of the Region of Interest (ROI), its feature extraction and classification of the detected/segmented lesion. CAD systems, being automated, expedite these subtasks, which otherwise are time-consuming and, hence, present the analysis results quickly. Consequently, the early detection of diseases and medical conditions becomes attainable. Moreover, they maintain consistency in all cases, no matter how long the patient list is, during the day. This is achieved because they

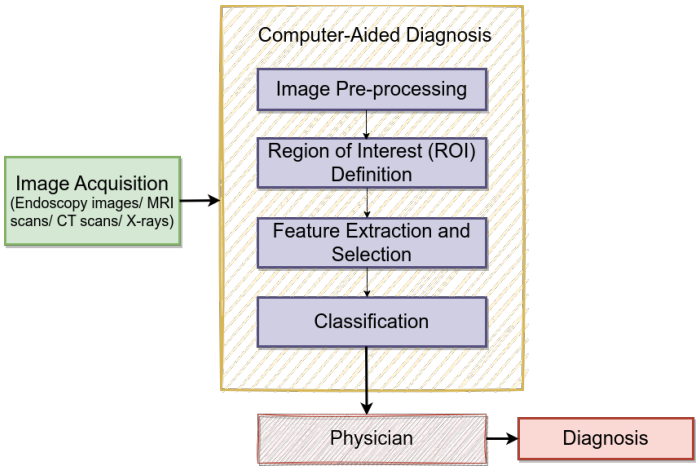


Figure 1.2: Block diagram of a general computer-aided diagnosis system.

do not experience fatigue or variations in performance and, thus, are less likely to encounter errors in the medical image interpretations.

Despite compelling results, some dependencies are also involved in this automated process. The quality of CAD outcomes significantly relies on the algorithms in the backend. Therefore, it is crucial to develop robust algorithms that can handle significant variations, including noisy data with corrupted images and can process a range of cases sampled from varied populations in terms of race, age, sex and demographic changes [4]. Besides algorithms, CAD's performance is influenced by the input data. It learns patterns from past case samples during training and uses them to associate with similar unknown future cases during testing. Hence, low-quality and under-sampled classes in the training data can result in biased predictions. Therefore, it can be inferred that high-quality input data representing different anomalies is an essential part of CAD systems. Considering the criticality of these concerns, our research encompasses medical data quality issues, unravelling their impact on clinical outcomes and developing a methodology to deal with low-quality data. Additionally, we provide solutions to data scarcity issues, subsequently addressing the challenges of class imbalance [5].

In the context of our above-mentioned objectives, we conduct research in the field of medical image analysis, particularly at the intersection of deep learning and medicine. Among the different imaging modalities discussed above, we mainly focus on *endoscopy and colonoscopy* data, i.e., the human *Gastrointestinal (GI) tract*. Endoscopy is a non-invasive procedure that examines the GI tract for any abnormalities. During this test, a thin tube mounted with a camera, known as an endoscope, is inserted through a natural opening such as the mouth. When the same process is performed for screening the colon,

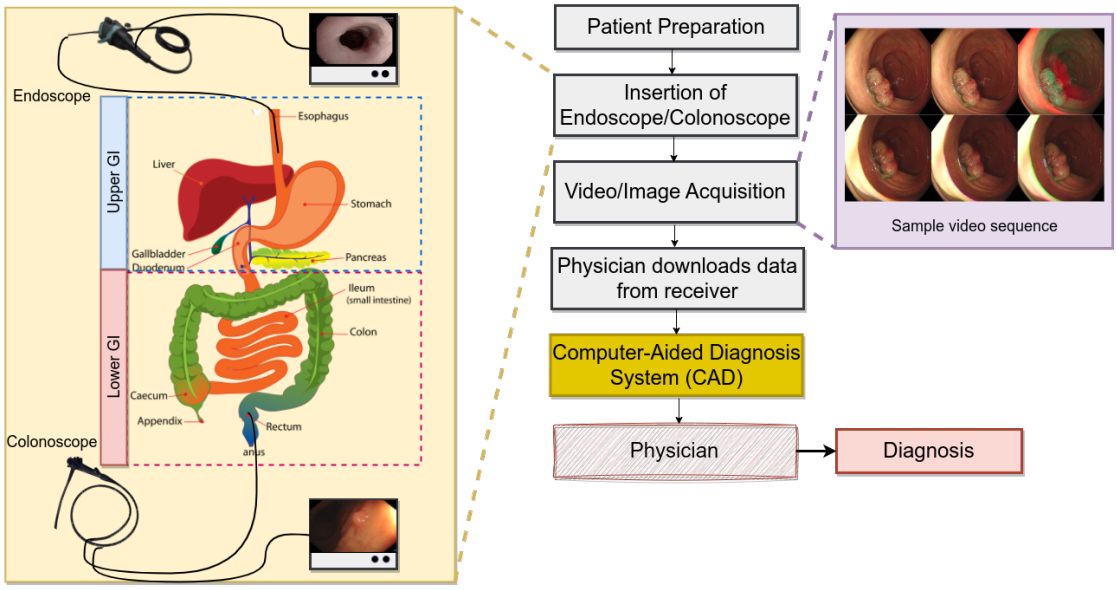


Figure 1.3: A visual description of the procedure and associated role of CAD.

and a colonoscope is inserted into the colon, the procedure is termed a colonoscopy. These medical examinations perform screening of the GI tract, which is a critical task to detect potential problems such as GI cancers (associated with the colon, stomach, pancreas, liver and esophagus), ulcerative colitis, colon polyps, inflammatory bowel disease, and gastroesophageal reflux disease. A visual description of the procedure and associated role of CAD are depicted in Fig. 1.3. According to a study [6], GI cancer cases are predicted to increase by 58%, and related deaths could show a 73% rise by 2040. In 2018, there were approximately 4.8 million new cases of GI cancer and 3.4 million deaths [6]. The five major types of GI cancers are colorectal (1.93 million cases; third most common cancer), pancreas (466,003 deaths; lowest survival rate), liver (905,677 cases), stomach (1.09 million cases), and esophagus (604,100 cases) [7]. Our research primarily encompasses CRC because it is the second leading cause of cancer-related mortalities and the third most common malignancy in the world, which accounts for about 10% of all cancer cases. CRC is marked by the growth of abnormal tissues called polyps. These polyps project out of the inner lining of the colon, and their type and characteristics are the determining factors for CRC and its level of risk. Timely identification of such biomarkers[3] and related diagnosis is of vital importance in reducing the rising CRC incidence and associated deaths.

[3]a measurable indicator of some biological state or condition.

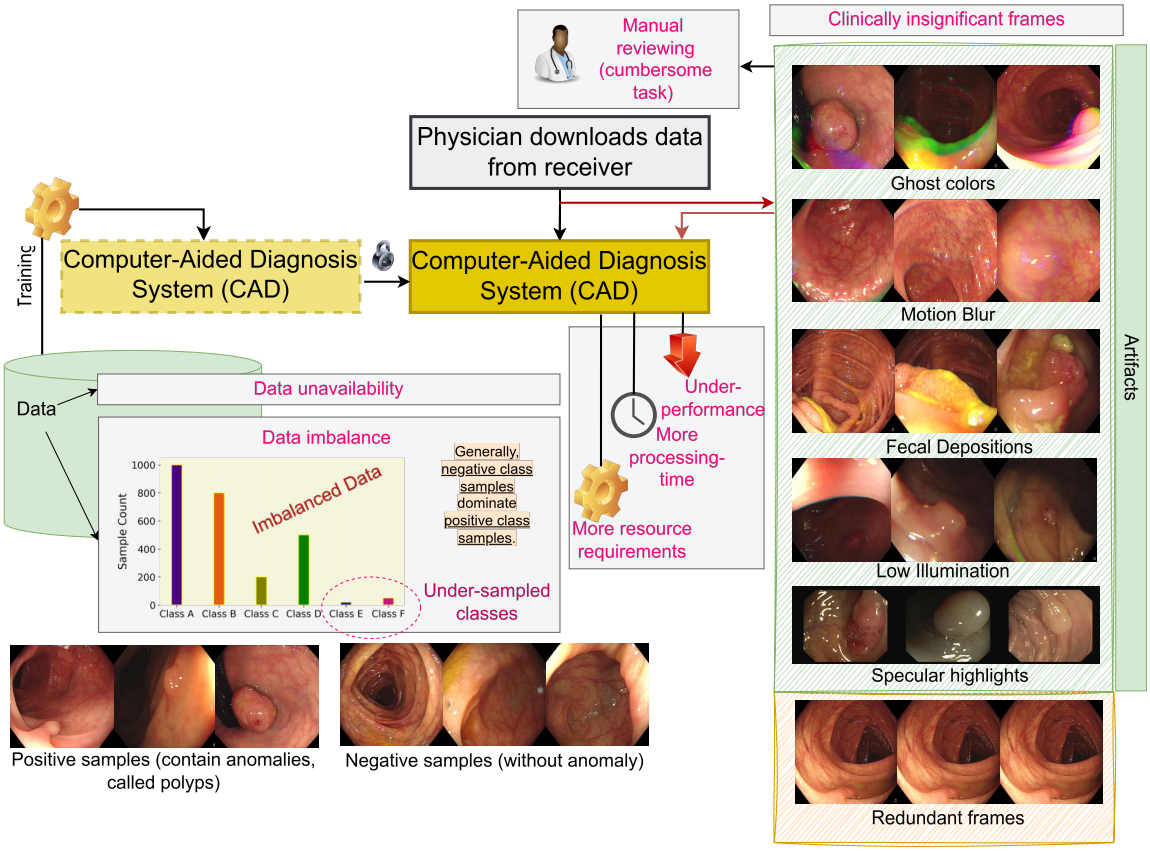


Figure 1.4: A flowchart describing the challenges associated with endoscopy/colonoscopy data review and clinical diagnosis using both manual and automated procedures.

1.2 Problem Description

This section describes the challenges of analyzing endoscopy data, particularly colonoscopy videos and images, in more detail. In addition, the objectives of this thesis and the underlying problem definitions are discussed. Further, the highlights of the thesis contributions are included. Fig. 1.4 provides a visual summary of the challenges. These are explained in detail below:

- **Large amount of insignificant/uninformative video frames:**
  - **Manual reviewing (cumbersome task):** The widespread acceptance of colonoscopy procedures as a gold standard for CRC screening is constrained by the massive amount of data recorded. During a standard colonoscopy, video encodings are done at 25 or 30 frames/sec. Such recordings impose a great challenge for clinicians to extract significant information expeditiously. If clinicians directly feed the video with unwanted frames to the automated detection sys-

tems, it would add to the computational cost. This is undesirable in case of limited resource availability. Hence, manual removal of frames is generally carried out by experts for better visualization and focused analysis. Such manual procedures are burdensome and induce human diagnostic errors.

  - **Presence of artifacts (underperformance):** During a colonoscopy, the arbitrary movements of the camera induce noise and motion blur in the recorded clips. Further, fecal depositions and food residues, even after bowel cleaning, inevitably introduce insignificant information [8]. It is estimated that around 25% frames of a colonoscopy video are low-quality. Some other artifacts responsible for low-quality frames include specular reflections, ghost colors, interlacing and low illumination. The presence of these artifacts could deviate the focus of a clinician or even lead to underperformance in automated techniques.
  - **Redundant frame processing:** The slow motion of the camera during colonoscopy results in substantial correlation among adjacent frames. Therefore, a large proportion of the video frames carry redundant information and, hence, do not add to clinical findings. Reviewing such frames imposes an unnecessary burden on clinicians and can negatively impact the CAD's performance by increasing computational costs and introducing noise in the input data.

- **Data unavailability:** As mentioned in Section 1.1, input data plays a pivotal role in influencing the accuracy of automated tools. Hence, training CAD systems with a significant amount of good-quality samples which can represent a considerably diverse set of abnormalities and target a large population is crucial. The lack of a diverse dataset impedes the fair comparison of deep learning models and reproducibility of experimental results, thus obstructing research studies.

  - **Multi-class dataset:** Most of the publicly available datasets [9–12] in the endoscopy domain are limited to a few classes, particularly the polyp class. Acquiring some anomalies during an endoscopy procedure could be difficult. This restricts the research to a few medical conditions. Therefore, the challenge is to curate a good-quality multi-class endoscopy image dataset that can help study a range of anomalies, improve deep learning models' performance, enable reproducibility, facilitate collaborations, and address bias. The term bias can be related to demography (considering under-represented groups), sex or hard-to-find anomalies.

  - **Barriers in dataset curation:** Medical data is generally sparsely available due to privacy concerns, legal restrictions and the time-consuming task of conducting manual annotations by multiple medical experts. Moreover, acquiring medical data can result in an imbalanced dataset as one medical condition may be rarer than another. For example, negative samples of colonoscopy images pertaining to non-polyp frames are easily accessible, whereas polyp frames are comparatively less acquired. Hence, the scarcity of large-scale datasets, especially that of the pathologically significant class, makes studying lesions and biomarkers challenging. This scarcity affects both the manual training of junior physicians and the automated training of diagnostic systems. Inadequate quality and distribution of data samples used for training automated tools produce error-prone outcomes. Therefore, it is crucial to train automated systems with ample amounts of good-quality samples representing the respective class with adequate information.

- **Polyp Detection, Segmentation and Classification:** Regardless of the surveillance using colonoscopy, the procedure suffers a high miss rate of polyps due to its dependency on the operator's ability. Moreover, the camouflage property of polyps makes them indistinguishable from the background, and their varied sizes and shapes make it difficult for clinicians to precisely identify polyps. Furthermore, flat or sessile polyps and polyps with size < 5 mm also contribute to the low detection rate.

  To overcome the challenges of manual polyp detection, several automated systems have been proposed [13–15]. However, a lingering gap in CAD performance still needs to be addressed. A few factors make these systems perform inadequately in some aspects: a) *Uncertain boundaries.* These models fail to accurately demarcate the polyp boundaries because these edges appear to be uncertain due to the polyp's high similarity with the background, b) *Imaging artifacts.* The inevitable introduction of artifacts in the colonoscopy images due to imaging conditions makes polyp regions fuzzy and vague, and this results in low performance of models in detecting polyps, and c) *Small-scale datasets.* Annotating the exact location and precise demarcation of polyps is a cumbersome task. This task requires domain expertise and an enormous amount of time, which limits the existing systems from relying on a small-scale dataset for learning the ability to detect and segment polyps.

  In the last few years, augmentation techniques like random rotation, flipping, cropping, and translation have become common to supplement the existing datasets with

more samples. However, such inclusion does not enhance the diversity or variations in the data samples. Moreover, prior works performed inferences on inconsistent splits, making it difficult to compare the performance of different methods in the literature, especially on small-sized datasets. Similarly, the challenges encountered in polyp classification are attributed to the low-quality data and significant class imbalance. The imbalance in some pathology classes results in biased outcomes, which is a critical issue. Due to the unavailability of good-quality image data with appropriate class distribution, only a few existing works [16, 17] focus on polyp classification.

1.3 Our Contributions

We present four contributions in this thesis. Our first contribution is focused on proposing solutions to reduce the manual effort of clinicians in reviewing a huge volume of colonoscopy data. The second contribution encompasses solutions to get rid of artifacts from the colonoscopy images and videos. As a third contribution, we curate a multi-class endoscopy dataset that is made publicly available to the research community. Lastly, in the fourth contribution, we propose a framework to generate synthetic data to overcome barriers encountered during real dataset curation. An overview of the contributions is illustrated in Figure 1.5.

Keyframe Extraction for Enhanced Polyp Detection and Localization: We propose a multi-stage methodology to extract keyframes to reduce the burdensome task of reviewing a large volume of colonoscopy data. Keyframes are the static summaries of the key-events captured during a patient's colonoscopy procedure. Prior works either focused on conventional methods to extract keyframes or ignored many of the crucial video attributes. The proposed framework aims to extract good-quality keyframes and considers the important video attributes of colonoscopy. These attributes include (a) low-quality, blurry frames with several artifacts, (b) high correlation in adjacent frames, and (c) cropped or distant polyp views. As a result, our framework selects keyframes that provide good-quality, non-redundant clinical information with multi-views and closer shots of polyps. We show that extracting keyframes helps improve the performance of state-of-the-art detection and localization deep learning models. It further reduces the computational cost and time requirements for analyzing a patient's video. Experimental results report that our approach removes 96.3% and 94.02% frames from the SUN Database [18] and the CVC-VideoClinicDB [19], respectively.

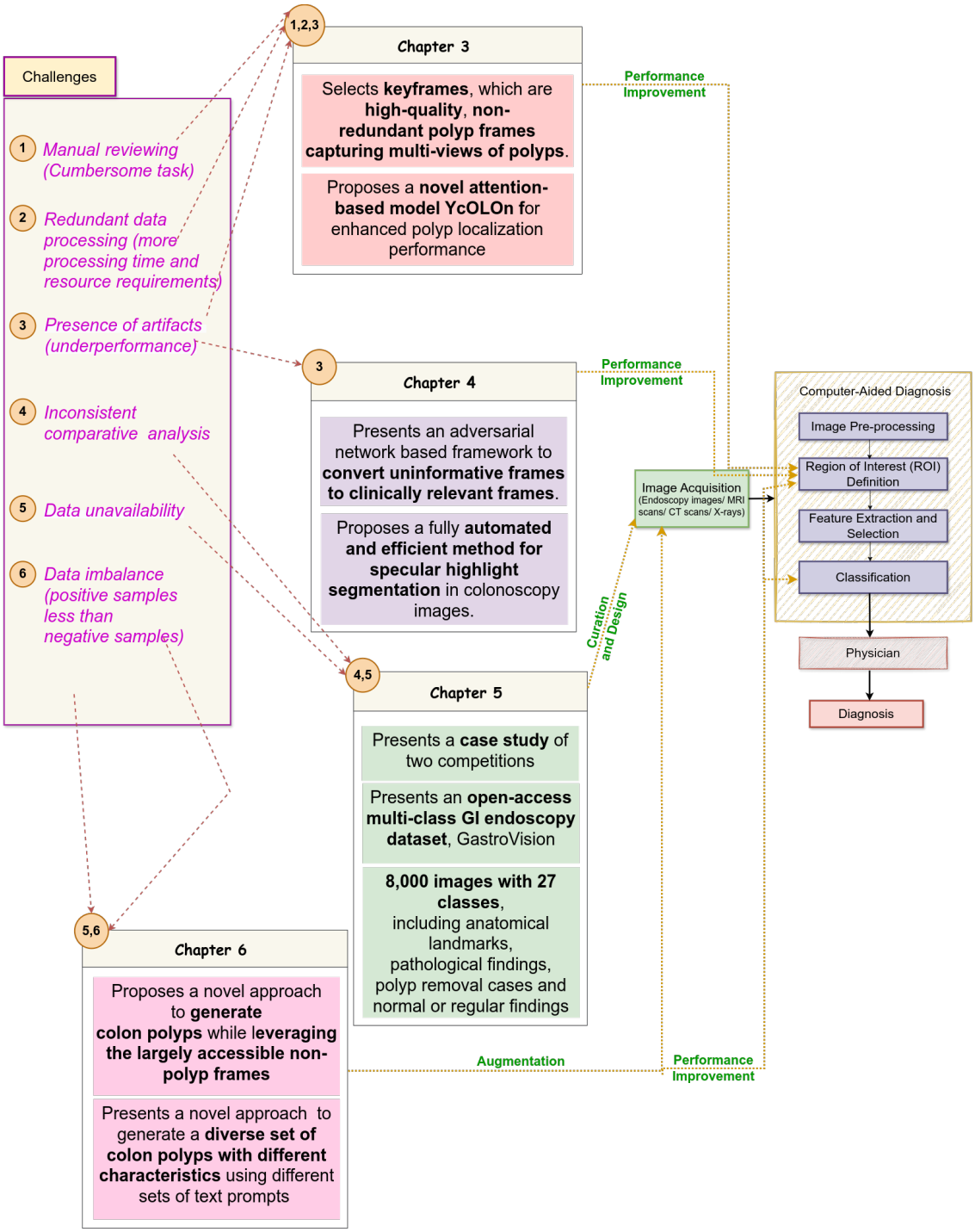


Figure 1.5: An overview of the thesis contributions.

In this framework, we also propose YcOLOn, a novel multi-scale attention-based model to localize polyps efficiently. We show that incorporating multi-scale attention blocks in the network helps enhance polyp localization performance by about 5.5% compared

to the baseline. Further, we develop a Graphical User Interface (GUI) application that assists in navigating through different stages of our framework. It facilitates the users with easy-to-use clickable options to load a video and obtain keyframes based on several options provided in the GUI.

Artifacts Removal: The predominant portion of the literature that focuses on artifact removal either deals only with specular highlights or completely removes unwanted blurry frames before utilizing them for downstream tasks. In the former case, the most common approach is threshold based which can easily fail in real scenarios. These methods can mainly be divided into three categories: (a) color distribution, (b) low rank & sparse decomposition, and (c) contrast based (explained in Section 2.3.2). The latter case, which removes insignificant frames, either follows a manual approach or completely removes such frames. The manual approach is a cumbersome task whereas the complete removal could sometimes result in data loss if the entire video carries corrupted frames due to inappropriate patient preparation or uncertain events.

To address these issues, in Sharma et al. [20], we propose an adversarial network based approach that translates uninformative frames to clinically relevant frames. This conversion helps extract the obscured details from blurry and corrupted frames. The technique is validated by comparing the detection/localization performance with and without the translation process. Preliminary results show that the detection performance improved with our approach. This study also provides insights into the effectiveness of generative models in artifacts removal, thus outlining pathways for future possibilities to explore such techniques.

In the adversarial approach, the whole frame is targeted for conversion. Such an approach is viable for artifacts like ghost colors, motion blur, low illumination, etc. However, specular highlights need a different approach, as they can be present in visually appealing, good-quality frames. These highlights are present in clusters in small regions and are the result of reflections from the watery inner surface of the colon. Therefore, to detect these highlights, in Sharma et al.[21], we propose an encoder-decoder based segmentation method. Considering the small-sized target in this case, we introduce the concept of Discrete Wavelet Transform (DWT) and use it to replace the standard pooling layers. This modification helps preserve the details pertaining to small regions of specular highlights. These layers are coupled with dilated convolutions for the same reason, as these convolutions increase the receptive

field while preserving input resolution. Further, we modify the standard low-frequency sub-band of DWT to help deal with overexposed regions. These overexposed regions are not categorized as specular highlights but can deviate an automated specularity detection algorithm from its intended task. An ablation study reports an increase in recall rate by about 1% with this modification while maintaining the precision, and embedding dilated convolutions further improved the precision by about 1%.

Case study, Dataset Curation and Design: Despite a lot of research in the area of polyp segmentation, there is a lack of fair comparison due to inconsistent splits followed in the literature. Moreover, most such studies fail to support their outcomes with transparency or interpretability validations. Further, some of the existing works are developed on private data and closed sources, restricting the reproducibility of results. Therefore, we present a case study on two competitions, namely, *"Medico Automatic Polyp Segmentation (Medico 2020)"* and *"MedAI: Transparency in Medical Image Segmentation (MedAI 2021)"* that performs a comprehensive analysis of several submissions received during the competition. The first competition involves two tasks: (a) polyp segmentation and (b) algorithm efficiency, and the second competition involves three tasks: (a) polyp segmentation, (b) instrument segmentation, and (c) transparency. This case study performs a fair comparison among different methods on standard metrics, which helps validate their feasibility in clinical settings. Also, the transparency task motivates for interpretable validations and open-source research work for reproducibility.

Considering the existing datasets and the datasets used in the case study, it can be inferred that prior datasets in the endoscopy domain comprise a limited number of classes. Also, most of these datasets focus on a single lesion, i.e. polyps and are small in size. Moreover, some of these datasets are accessible on request, which induces unnecessary delays. Therefore, we develop *GastroVision*, an open-access multi-class Gastrointestinal (GI) dataset that consists of 8000 images from 27 different classes. Our dataset not only covers pathological abnormalities but also includes anatomical landmarks, normal findings and polyp removal cases. Most of the samples are acquired through White Light Imaging (WLI), and a few images belong to Narrow Band Imaging (NBI). The findings in *GastroVision* can broadly be categorized as upper GI and lower GI tract. A series of experiments are performed using six baseline deep learning models. Our results are reported on standard multi-class classification metrics that allow the research community to build more robust solutions for better

outcomes. The findings in our dataset are annotated by experts and are acquired from two different centers using standard equipment from Olympus and Pentax. The curation process complies with the ethical and privacy norms. The dataset is publicly available at https://osf.io/84e7f/.

Synthetic Dataset Creation: In recent years, the field of generative Artificial Intelligence (AI) has gained prominence in the medical domain. The literature in the colonoscopy domain covers many Generative Adversarial Network (GAN) based techniques to generate synthetic images. However, these approaches show limited success and suffer from convergence instability. More recently, diffusion models have gained popularity for their realistic image generation. Considering these remarkable outcomes, in Sharma et al. [22], we propose *ControlPolypNet*, a novel diffusion based network that converts non-polyp images into polyp images using a custom binary mask. Our method leverages easily accessible non-polyp frames to obtain hard-to-find polyp frames. Additionally, this generation process is controlled using a novel user-configurable input control map. This control map uses a non-polyp image overlapped with a polyp mask. The polyp mask is user-defined, and therefore, the polyp location, size and shape can be easily controlled.

Limited works in the literature utilized diffusion models to obtain polyp images; however, these works use simple binary masks as input. Unlike such approaches, we leverage non-polyp frames overlapped with binary masks to preserve other endoluminal elements. This approach reduces the probability of obtaining unwanted structures or noise in the background/endoluminal scene. Further, we introduce a detector in our framework to ensure pathologically relevant data at the end. We validate the quality and clinical significance of our synthetic data by using it to augment two publicly available datasets in the polyp segmentation task. Experimental results show that an average increase of 1.3% to 6.84% in the Jaccard Index (JI) is obtained using our technique. Additionally, our above approach provides a pathway to address the issue of class imbalance as one undersampled class can be expanded using synthetic data. In our case, the positive class (polyp) has relatively fewer samples than the negative class (non-polyp). We utilized negative class samples to produce positive class samples, thus expanding the undersampled set.

Furthermore, we explore another control mechanism based on text prompts, which has not yet been explored in the literature. While utilizing this control mechanism,

in Sharma et al. [23], we propose *PathoPolyp-Diff*, a novel diffusion based model that generates text-controlled synthetic images. The model is capable of generating colonic polyps that cover a wide range of classes in terms of imaging modalities (NBI/WLI), pathology (adenomatous/hyperplastic) and quality. These medical terminologies are discussed in detail in Section 2.2. In this work, we also propose a cross-class label learning methodology that allows the model to learn characteristics from other classes. This approach reduces the need for the burdensome task of data annotation and provides flexibility in generating diverse sets of polyp types. We validate the performance of our technique by augmenting a publicly available dataset with synthetic data obtained using different text prompts. The results report an increase of 6% in the Balanced Accuracy (BA) on a downstream task that classifies pathology classes combined with different imaging modalities.

1.4 Thesis Outline

The thesis comprises seven chapters and is organized as follows:

In ***Chapter 2***, we first describe essential terminologies and background, introducing the basic concepts associated with the GI tract. We further provide a detailed review of existing methods followed by an overview of evaluation metrics used in the thesis.

In ***Chapter 3***, we present a multi-stage framework for keyframe extraction from colonoscopy videos. It includes a patient-wise analysis and also shows how it becomes difficult to select keyframes for some patients with an unacceptable amount of uninformative frames. This limitation serves as a motivation for the next chapter.

In ***Chapter 4***, we present two solutions for artifacts removal from colonoscopy images and videos. The first solution is based on an adversarial network and focuses on artifacts such as ghost colors, low illumination, motion blur, interlacing, and fecal depositions. The second solution mainly aims to perform the segmentation of specular highlights.

In ***Chapter 5***, we present a case study that includes a comprehensive summary and analysis of two competitions, namely, *"Medico Automatic Polyp Segmentation (Medico 2020)"* and *"MedAI: Transparency in Medical Image Segmentation (MedAI 2021)"*. Each competition involves submissions from 17 different teams. The chapter showcases the analysis of each contribution, highlights the best-performing methods and discusses their viability in clinical settings.

Further, we present a multi-class open-access GI dataset with 8000 images and 27

classes. The chapter provides a detailed description of the dataset, including class-wise sample distribution, sample-wise resolution, and a broader set to which the 27 classes belong. It also presents experimental results on six baseline deep learning models.

In ***Chapter 6***, we present ControlPolypNet and PathoPolyp-Diff, two diffusion based models for generating synthetic colonoscopy images. The chapter demonstrates the effectiveness of these generative methods and the quality of synthetic samples by performing downstream tasks of polyp segmentation and classification.

Finally, in ***Chapter 7***, we conclude and discuss future research work.

2

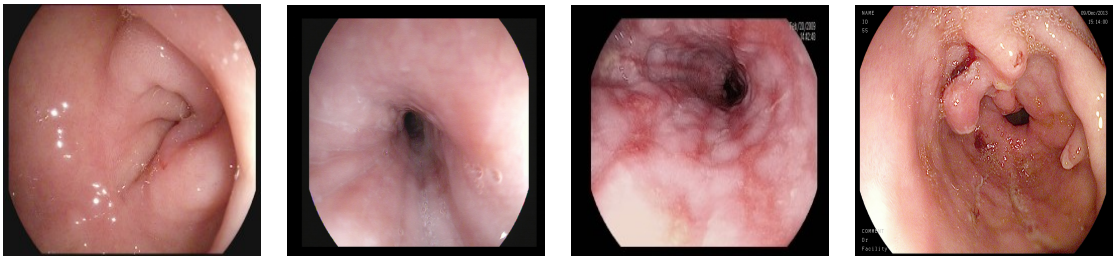
Background and Literature Survey

This chapter provides the background necessary to understand the thesis better. We begin by describing essential terminologies and concepts, followed by a discussion on conventional modelling approaches and standard evaluation metrics.

2.1 Gastrointestinal Tract and its Examination

The Gastrointestinal (GI) tract, also known as the digestive tract, comprises a series of organs involved in the movement of food and liquid using a process called peristalsis. These organs help to break food into smaller chunks, which allows the essential nutrients to get absorbed by the body, and finally, the waste residue is expelled. The complete process is associated with both upper and lower GI tracts. The upper GI tract refers to the esophagus, the stomach and the duodenum (first part of the small intestine), whereas the lower GI tract comprises the small intestine and the large intestine. Each organ has a different function and manifests different symptoms in case of any disorder, but they are connected together via a tubular pathway. During healthcare procedures, identifying the location of the affected organ is essential for understanding anatomy, creating precise surgical plans and determining optimal treatment strategies. Professionals track anatomical landmarks, such as the z-line, pylorus and cecum, which serve as reference points to locate the site of interest. Some of the anatomical and pathological findings (disorders) and the related examination techniques are discussed below.

Upper GI tract: Fig. 2.1 shows sample images of different anatomical and pathological findings of upper GI.

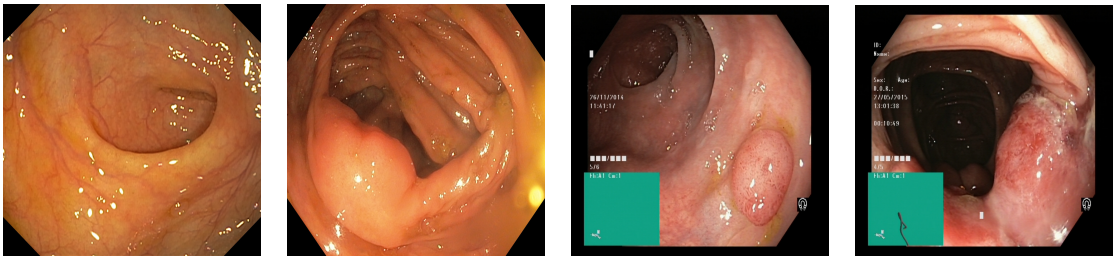


(a) (b) (c) (d)

Figure 2.1: Upper GI tract: (a) Pylorus, (b) Normal esophagus, (c) Esophagitis, and (d) Gastric polyps.

- *Anatomical Landmarks:* The three main anatomical landmarks include the *normal z-line*, the *pylorus* and *retroflux stomach*. The *normal z-line* represents the transition between the squamous esophageal mucosa and columnar mucosa lining of the stomach. *Pylorus* demarcates the end part of the stomach and the beginning region of the small intestine called the duodenum. *Retroflex stomach* denotes the retroflexion of the endoscope that involves turning the camera to visualize the upper part of the stomach.
- *Pathological Findings:* Pathological findings signify the abnormal modifications in the tissues of mucosa indicative of a disease. Some of such conditions are *esophagitis*, *Barrett's esophagus*, *ulcers*, *gastric polyps* and *cancer*. Early detection of the pathological findings could assist in providing treatment before the development of cancer.
- *Examinations:* Upper endoscopy or EGD (esophagogastroduodenoscopy) and the upper GI series are the common procedures followed to examine upper GI for any abnormality. The upper GI series is a radiographic test that uses fluoroscopy, X-ray or CT scans to examine esophagus, stomach and duodenum. Fluoroscopy, a special form of X-ray, sends a continuous radiation beam for a few seconds to capture a video and, hence, can be considered an "X-ray movie". During these procedures, no tool enters the patient's body. Contrarily, an endoscopy requires the insertion of an endoscope through esophagus. An endoscope is a thin, long, tube-like instrument with a camera mounted on the tip. The camera captures video of the patient's affected region and directs them to a monitor screen. Compared to radiology-based tests, it provides more visual details.

Lower GI tract: Fig. 2.2 shows sample images of different anatomical and pathological findings of lower GI.



(a) (b) (c) (d)

Figure 2.2: Lower GI tract: (a) Cecum, (b) Ileocecal valve, (c) Colon polyp, and (d) CRC

- *Anatomical Landmarks:* Some examples of lower GI anatomical landmarks are *Cecum*, *Ileocecal valve*, and *Terminal ileum*. *Cecum* is a large pouch that is present at the proximity of the large intestine. *Ileocecal valve* is a valve forming junction between *Cecum* and the *Terminal ileum*, which maintains flow between the two anatomical landmarks.
- *Pathological Findings:* The pathological conditions that might occur in the lower GI include *Angiectasia*, *Colon polyps*, and CRC. Among the lower GI pathological findings, CRC is a severe condition with an estimated high rate of 1.6 million mortalities per year, expected by 2040 [24]. In 2020, it contributed to 10% and 9.4% of new cancer cases and cancer-related deaths, respectively. With such critical statistics and numbers increasing at an alarming rate, it becomes crucial to address the issue in the early stages to prevent any benign growth from developing into CRC over time. A more detailed discussion of CRC is given in the subsequent sections.
- *Examinations:* Similar to upper GI, a patient for lower GI screening is suggested to undergo either a colonoscopy or a lower GI series. The lower GI series is a radiology-based test, among which Barium X-ray is a commonly performed test. During this process, the large intestine is filled with Barium liquid. It is an X-ray absorber and appears white on the X-ray films, which helps enhance the visibility of different GI characteristics. Another preferred test, called endoscopy, which includes sigmoidoscopy and colonoscopy, is a widely adopted procedure for examining the lower GI. Although sigmoidoscopy and colonoscopy are both used for CRC screening, the former looks for any abnormality in the lower part of the colon, whereas the latter tests the complete large intestine. Besides endoscopy, other recommended tests are stool tests, virtual colonoscopy (CT colonography), Fecal Occult Blood Test (FOBT) and Fecal Immunochemical Test (FIT).

2.2 Colorectal Cancer (CRC)

CRC is the third most common malignancy and the second leading cause of cancer-related mortalities. It begins as an abnormal growth on the inner lining of the colon, forming clumps of cells called polyps. These polyps are generally non-cancerous but have the potential to turn into cancer over time. In the earlier stages, CRC does not show any symptoms, and therefore, it becomes critical to undergo regular screening tests to detect any precursor to avoid severe CRC conditions. In the later stages, rectal bleeding, change in stool color, and pain are some symptoms that might be observed.

How likely is it to develop CRC?

The statistics[1] released by *American Cancer Society* report disparities in the chances of someone developing CRC based on demographic and geographic variations. It mentions that approximately 4.4% of men and 4.1% of women are likely to be diagnosed with CRC in their lifetime. Men are at 30% more risk of CRC incidence compared to women. This variation could be attributed to exposure to risk factors such as cigarette smoking. Besides gender, it also varies across age groups. Although earlier studies state that older age groups are susceptible to CRC risk, recent statistics show that the incidence rate of CRC among people aged 50 and older is declining, whereas it is increasing among younger age groups. Consequently, the median age of CRC diagnosis has shifted from 72 years (in the 2000s) to 66 (in the current scenario). Such disparities are further related to ethnic and racial differences. The non-Hispanic blacks (hereafter, blacks) are the most affected by the CRC incidence and death rates, followed by American Indians and Alaska Natives (AIANs), and the rate is lowest in Asians/ Pacific Islanders. Across geographic boundaries, socioeconomic factors, dietary patterns and access to high-quality healthcare facilities are some of the reasons for variations in CRC incidence rates. Currently, the poor countries observe 30% to 40% higher CRC-related deaths compared to rich countries. The geographical factors are more influential than the racial factors, as blacks and whites tend to show similar risks when sharing the same geographic boundaries.

[1] Colorectal Cancer Facts & Figures 2020-2022. Link: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/colorectal-cancer-facts-and-figures/colorectal-cancer-facts-and-figures-2020-2022.pdf

2.2.1 Polyps and their Categorization

A polyp is a small abnormal growth in the form of excess tissues that project on the inner lining of the colon. The majority of such occurrence is non-cancerous, but some of the polyps pose the risk of cancer over time. Due to this reason, a physician looks for such polyps during colonoscopy procedures to monitor any progression and analyze their stage and associated treatments. For appropriate analysis and to plan treatment strategies, different polyp characteristics are taken into account. Polyps can be categorized based on multiple factors such as size, shape (Paris classification), texture and surface patterns (NICE classification) and pathological condition.

Paris classification

The Paris classification [25] is a standard polyp classification system defined based on the polyp shape. A broad categorization covers two types, namely, *polypoid (type 0-I)* and *non-polypoid (type 0-II and type 0-III)*. The polypoid class involve polyps with a protrusion above the mucosa, whereas polyps under the non-polypoid class do not protrude. *Type 0-I* can be divided into *pedunculated (type 0-Ip)*, *semi-pedunculated (type 0-Isp)* or *sessile (type 0-Is)* polyps. Similarly, *type 0-II* also have subcategories, which include *slightly elevated (type 0-IIa)*, *completely flat (type 0-IIb)*, and *depressed (type IIc)*. Lastly, the second main type of non-polypoid class includes *excavated (type 0-III)* form of polyps.

There are some distinguishing criteria for each category. For example, the *Type 0-Ip* is different from the *Type 0-Is* in terms of the base and top diameter. In the former, the base is narrow, whereas in the latter case, the polyp possesses the same diameter for both the base and the top. In the non-polypoid subclasses, the distinguishing criterion is the amount of elevation the polyp has compared to its surrounding mucosa. The type 0-IIa, being similar to type 0-Is polyps, needs some further investigation to decide the actual type. Some non-polypoid category polyps spread laterally (>10 mm) without increasing protrusion above the mucosa. Such polyps are termed as *Lateral spreading tumors*. A flowchart illustrating different shapes based on Paris classification is shown in Fig. 2.3.

NICE classification

NBI International Colorectal Endoscopic (NICE) classification considers the vascular and surface patterns to divide polyps into three categories, i.e., *hyperplastic (type 1)*, *adenoma (type 2)* and *deep submucosal invasive cancer (type 3)*. *Type 1* polyps have lighter or similar

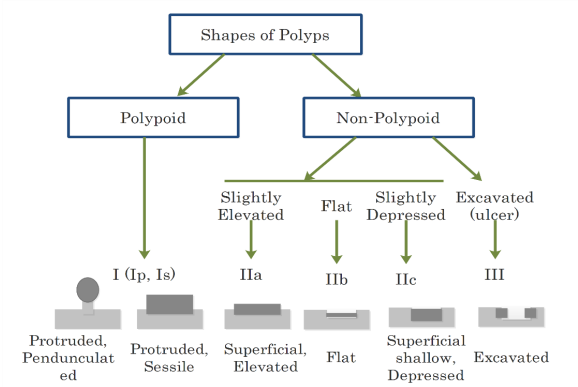


Figure 2.3: A flowchart depicting Paris classification

(a) (b) (c)

Figure 2.4: NICE classification (a) Type 1, (b) Type 2, and (c) Type 3. Image Source: https://www.endoscopy-campus.com/en/classifications/polyp-classification-nice/

color as that of adjacent mucosa, have either no or some isolated lacy vessels and might have uniform dark or white spots. *Type 2* polyps appear more brown compared to the background and have brown vessels surrounded with white structures. They also have tubular, branched or oval white structures over the surface. *Type 3* polyps have a brown to dark brown color relative to the background and might have discontinued or broken vessel structure. Their surface pattern is either amorphous or absent. Some sample images are shown in Fig. 2.4.

Pathological classification

Polyps can be broadly classified as neoplastic and non-neoplastic. Neoplastic polyps include *adenomatous* polyps (further classified as *Tubular adenomas*, *Villous adenomas*, and *Tubulovillous adenomas*) and *sessile serrated*. Non-neoplastic polyps can be sub-categorized as *hyperplastic* polyps, *inflammatory* polyps, and *hamartomatous* polyps. Among these classes, *adenoma* and *hyperplastic* are the most common terms used to represent potentially cancerous and non-cancerous lesions, respectively. A different aspect of polyp classification introduces the term *Serrated* polyps, which include *hyperplastic* polyps, *sessile serrated* le-

sions, *traditional serrated adenomas*, and *unclassified serrated adenomas*. *Serrated* polyps have saw-toothed appearance at microscopic level and contribute to 25% of CRC cases[2].

Tubular adenomas, being the most common type in the adenoma class, accounts for 80% of the *adenomatous* polyps. Only a small proportion of such polyps tend to convert into cancer; however, CRC is most likely reported to start from this category of polyp. Therefore, *Adenoma Detection Rate (ADR)* plays a significant role in determining the effectiveness of CRC preventive methods. *Villous adenomas* are more probable to become cancerous, but they contribute to 5%-15% of all *adenomatous* polyps. A similar ratio is estimated for *Tubulovillous adenomas*. The findings and pathologic information about *sessile serrated* has evolved drastically over the past few years. Initially, this category of polyps was defined under *hyperplastic* class. However, in recent years, they were found to carry some malignancy risk [26]. Here, the term "sessile" denotes the slightly elevated or flat shape of the polyp. The incidence rate of *traditional serrated adenomas* is very low and *unclassified serrated adenomas* covers those polyps which present a mixed sign, i.e., resemble adenoma, but appear similar to sessile and serrated polyps. Contrary to the above neoplastic examples, *Inflammatory* and *hamartomatous* polyps are much less probable to become cancerous.

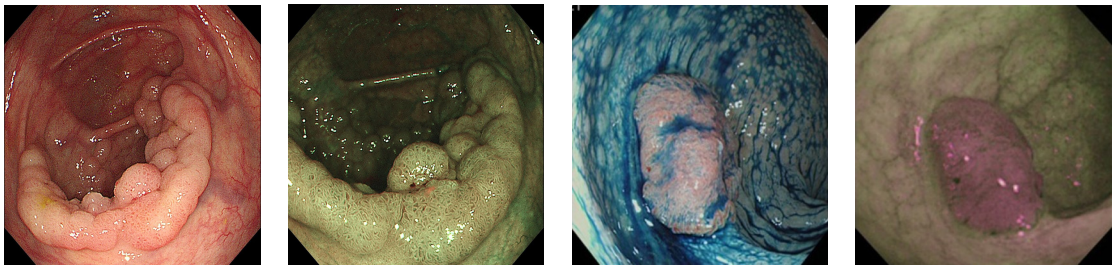
Size-based classification

Polyp size is considered an important factor in determining the cancer risk. According to a study [27], cancer rates are correlated with polyp size. It was observed that out of all polyps with adenomatous features, 25% were in the range of 1-9 mm size and 75% were either 10 mm or larger than 10 mm. Based on the size, a polyp is categorized into three divisions, namely, *diminutive* ($\leq$ *5 mm*), *small* ($\geq$ *6 mm,* $\leq$ *9 mm*), and *large* ($\geq$ *10 mm*). While *diminutive* and *small* polyps carry low CRC risk, *large* polyps if found with adenoma histology, are recommended for short interval follow-ups (generally 1 year) for surveillance colonoscopy [28].

2.2.2 Colonoscopy and CRC Treatment

Colonoscopy is a minimally invasive procedure to examine the colon and is considered the gold standard for CRC screening. It is an endoscopy during which a colonoscope is passed to the colon through the anus and rectum. A colonoscope is a long flexible tube with a light and a camera at one end which streams the video from the inner regions of the

[2] https://my.clevelandclinic.org/health/diseases/17462-serrated-polyps



(a) (b) (c) (d)

Figure 2.5: Colon polyp under: (a) WLI, (b) NBI, (c) Dye-spraying, and (d) AFI. Image Sources: [29, 30]

colon on a monitor. Thereafter, the captured colon video is examined by a physician for any irritated or swollen tissues, polyps or cancerous lesions. During a colonoscopy, the colonoscope tube also helps clean the inner colon lining through a water jet and injects air to improve visibility. While advancing the colonoscope through the colon, in addition to video capturing, the physician may remove any abnormal tissue or can take any sample for biopsy.

Imaging technologies

As discussed in Section 2.2.1, CRC polyps are characterized by three-dimensional surface topographic features, which make them distinguishable from the surrounding mucosa. Such differentiating polyp-specific features are significant and are essential to capture during colonoscopy. To support this crucial requirement, various optical imaging technologies have been adopted in colonoscopy. The most widely used imaging technology is White Light Imaging (WLI) endoscopy because it is cost-effective, medical professionals are trained in and accustomed to using WLI endoscopy, and it is readily available in most medical facilities. While the standard WLI endoscopy can enhance coloration, it fails to capture topographic contrast, such as polyp elevation and pit patterns that increase the lesion miss rate. To alleviate this issue, other imaging technologies such as High-definition White Light Imaging (HD-WLI), Narrow Band Imaging (NBI), Dye-based chromoendoscopy, Blue Light Imaging (BLI) and Linked Color Imaging (LCI), I-scan digital contrast and Flexible Spectral Imaging Color Enhancement (FICE), and Autofluorescence Imaging (AFI) were introduced [29]. Some sample images captured using different modalities are shown in Fig. 2.5.

HD-WLI colonoscopy provides higher resolution images and comparatively more image count per second compared to standard WLI. It has been reported that a 4.5%-12.6% increase in ADR can be achieved by opting for HD-WLI over standard WLI [31]. In dye-

based chromoendoscopy, contrast dyes are sprayed over the mucosa to improve the visibility of polyp surface topography. The other remaining imaging techniques, like NBI, LCI, and I-scan, are categorized under virtual chromoendoscopy, among which NBI is the most adopted technique. NBI uses electronically activated filters to limit the wavelengths of red, green and blue lights, which helps accentuate the superficial mucosa and vascular details. Similarly, the I-scan enhances the mucosal vasculature and topographic details but achieves this using three different modes, namely, surface, contrast and tone enhancements. With the same objective, FICE is introduced, in which light bandwidth is narrowed down using a computed spectral estimation system. Unlike the above-mentioned techniques, BLI and LCI utilize laser light sources to highlight the abnormalities and provide bright images [31].

CRC Treatment

The treatment methods have evolved rapidly over time with advanced techniques in healthcare. Moreover, to decide on a specific treatment, various criteria such as tumor type, location and patient characteristics are considered. In most cases, people with CRC undergo surgery. However, the specific method followed depends on the cancer stage. For example, if the CRC is in a localized stage, i.e., cancer has penetrated the colon wall but not completely, the resection of the affected region is performed, and some normal tissues from the surrounding region and nearby lymph nodes are removed. Contrarily, in a regional stage, when cancer has infected the nearby lymph nodes and has penetrated into the colon wall, then apart from the surgical resections, chemotherapy is usually recommended.

Therapeutic interventions: A polypectomy is a minimally invasive surgical procedure and the commonly adopted therapeutic intervention to remove polyps. During the process, a clinician uses a forceps or a snare to snip off the polyp. The forceps-grasping tool can be used to completely remove small polyps. The other instrument, the snare, comes with a wire loop. This loop can be placed in a way that it tightens around the polyp's base so that it cuts the polyp region out. This procedure can be carried out using one of the suitable treatments from the three options, which are *Hot Snare Polypectomy (HSP)*, *Cold Snare Polypectomy (CSP)*, and *cold or hot forceps* polypectomy. In HSP, a heated snare is used, which might be followed by electrocautery that helps burn the residue of any infected tissue. Similarly, CSP involves cold snare and is adopted for diminutive polyps. The last option is used for small polyps, which are pulled loose and then removed.

2.3 CAD and CRC Diagnosis: A Review on Methods and Metrics

This section presents a review of the different tasks undertaken by an ideal CAD system to perform CRC diagnosis. The sequence of these tasks followed in this section can be given by: *keyframe selection –> artifacts removal –> polyp detection and localization –> polyp segmentation –> polyp classification*. We also discuss existing related works in the literature, their proposed algorithms in brief and their shortcomings. In addition, we provide a detailed description of existing datasets to highlight the need for our proposed multi-class dataset. Also, we mention prior works focused on expanding real datasets with synthetic data and describe their drawbacks. Further, we include a description of the standard metrics used for assessing the different algorithms.

2.3.1 Keyframe Selection

A keyframe selection process allows to select keyframes (important frames) from colonoscopy videos that can efficiently represent the clinically relevant information captured in the video streams. For example, given a colonoscopy video (V) represented as a sequence of $\mathfrak{m}$ frames ($[f_1, f_2, f_3, ..., f_{\mathfrak{m}}]$), the goal of keyframe selection is to find a subset of frames that are clinically significant and representative. We denote keyframes of the original video V as $V_{key} = [f_{i_1}, f_{i_2}, f_{i_3}, ..., f_{i_k}]$, where $k << \mathfrak{m}$ and $V_{key} \subset V$. This procedure must take into account important attributes of a colonoscopy video. These attributes are (a) low-quality frames containing various artifacts (detailed explanation in Section 2.3.2), (b) a large number of redundant frames, and (c) many frames with a cropped or distant view of the polyp.

Related Work

In the colonoscopy domain, very few works are fully dedicated to keyframe selection. However, a common procedure generally followed for the pre-processing of colonoscopy videos involves quality assessment, an initial step of keyframe selection. For example, Park *et al.* [32] adopted two measures for frame quality: Shannon's entropy and a range filter. In a similar domain of Wireless Capsule Endoscopy (WCE), Yuan and Meng [33] performed hierarchical keyframe extraction. They utilized the entropy change to find the local maximum to create sub-clots. It is then followed by an affinity propagation clustering approach to select representative frames from each sub-clot. Oh *et al.* [34] proposed edge-based

and clustering-based techniques to classify frames as informative/uninformative. They also proposed a specularity detection method and used it further to improve informative frame classification. Arnold *et al.* [35] carried out wavelet analysis to detect indistinct frames. A recent work focusing on video summarization is proposed by Li *et al.* [36]. Some clustering-based approaches are presented in [37, 38]. However, clustering-based methods are not suitable in noisy environments. Colonoscopy frames are generally susceptible to noise. Also, redundant frames are captured during the colonoscopy, making clustering methods perform poorly. Saliency maps for finding keyframes of videos were presented in [39]. Mendi *et al.* [40] adopted a color histogram comparison-based method, which compared the color histogram of successive frames in a video. Keyframes were selected using k-means and PCA whenever a significant change was observed. However, this model does not fit into colonoscopy videos as most frames have similar color information. Recently, dictionary learning based approaches have been proposed for video summarization [41]. Sasmal *et al.* [42] utilized conventional methods, including image moments, Sobel operators, Oriented FAST, and Rotated BRIEF methods. They also incorporated depth information obtained using transfer learning.

Some other medical domains which focus on similar issues include laparoscopy, gastroscopy, and hysteroscopy. Ma *et al.* [43] extracted keyframes from laparoscopy videos. They used deep features and formulated video summarization as a diverse and weighted dictionary selection model. Loukas *et al.* [44] used an objectness model to segment the laparoscopy videos; and then the representative frames were selected based on the highest state-conditional probability. Xu *et al.* [45] considered eliminating uninformative frames and skipped anatomy detection results in such frames. The approach processes all frames simultaneously for detection and informative/uninformative classification. Hence, it does not reduce the processing time. Such related literature is included in a review study by Jin *et al.* [46], which covers disease-related and non-disease-related work. The non-disease-related papers are based on the frame quality for better lesion detection. In [47], a gastroscopic video summarization technique based on a dictionary learning approach is proposed. Ejaz *et al.* [39] proposed a visual attention-driven framework to perform video summarization in the field of hysteroscopy.

Limitations: Most of the above approaches use hand-crafted features and do not incorporate many video attributes leading to inefficacy in significant frame selection. Some of the methods do not consider noisy frames with ghost colors, which is a common issue in most endoscopy video recordings.

(a) (b) (c) (d) (e)

Figure 2.6: (a) Specular highlights (marked by blue bounding boxes over polyp), (b) Low-illumination, (c) Ghost colors, (d) Motion blur, and (e) Fecal depositions.

2.3.2 Artifacts Removal

The image-capturing environment and procedure (movements of colonoscope and watery surface inside the colon), during a colonoscopy, introduces several artefacts in the video stream, thus degrading the image quality. These artifacts can be in the form of low illumination, specular highlights, interlacing, ghost colors and motion blur. Some sample images are shown in Fig. 2.6. Controllable up to a certain extent with device settings, these artefacts are nearly inevitable. Such artifacts can deviate a computational model's intended task, reducing its performance.

Related Work

Most of the existing artifacts removal approaches are focused on specular highlight segmentation. Removal of other artifacts is an under-studied problem. The literature related to specular highlight segmentation is summarised in this section in three categories: (a) Color distribution, (b) Low rank & sparse decomposition, and (c) Contrast based methods.

Color distribution properties of colonoscopy images have been explored by [48, 49] to get a more comprehensible view of the highlights. Stehle *et al.* [48] has examined two color spaces. One color space can visualise both the textural and specular edges, while the other one is sensitive only to the textural edges. The difference between the two obtained images, along with saturation and brightness value tests, has been utilized to detect specular reflections. Gross *et al.* [49] converted images into HSV color space because specular highlights are characterised by low value and high saturation. Akbari *et al.* [50] proposed a method in which statistical features are extracted from each channel of RGB and HSV color spaces. Based on these features, a non-linear SVM selects one of the color spaces. RGB detection method employs the voting scheme, and the other one uses thresholding that is adaptive to the image statistics. Figueiredo *et al.* [51] computed the summation of respective intensities in the three channels of RGB color space and then laid a threshold over it. Yu *et al.* [52]

proposed an adaptive specular highlight detection method based on the ratio of the red channel to the green and blue channels. They considered that in non-specular regions, the value of the red channel is likely to be higher than the other two channels. On the contrary, all three channels have similar and saturated values in the specular regions.

Limitations: In a colonoscopy, the amount of light incident determines the intensity of bright spots. Consequently, these specular highlights seem to be darker in less illuminated regions. Hence, intensity thresholding could not result in acceptable detection outcomes. Moreover, these methods [48–52] are sensitive to noise and are highly dependent on the selection of thresholds. Such approaches are not robust and can lead to under or over-segmentation.

Low rank & sparse decomposition methods decompose the given image into low-rank and sparse results. Specular reflections are represented by the sparse results as these methods consider highlight pixels as noise. Li *et al.* [53] converted RGB images into HSV color space and then applied adaptive thresholding on the intensity and saturation values. The sparse results and the output obtained after applying adaptive thresholding are tested for similarity. After computing similarity measurements, they are used for performing iterative optimization in adaptive RPCA decomposition. This gives the final adaptive sparse results.

Limitations: The above method fails when there is too much specularity present in the image as it does not treat specularity as a sparse part in such cases. Li *et al.* [53] considered overexposed areas as specular regions, contrary to the ground truth provided by the experts. This method only focuses on the absolute highlight pixels and cannot detect relative highlight pixels.

Contrast-based methods search for high contrast regions and pixels corresponding to high intensity. Park *et al.* [32] detected saturated as well as high contrast areas. They utilized adaptive thresholding on the image intensity histogram to detect saturated regions. To search for small bright regions, an open top-hat filter followed by a reconstruction operation using a size-5 disk-shaped filter was adopted. In [54], it has been assumed that the specular reflections' intensity is more than its neighboring non-specular regions. The images and their corresponding medians are subtracted, and thresholding is applied to detect the bright spots. Sanchez *et al.* [55] proposed a method that focuses on high-contrast regions. The detected regions are fed to an SVM classifier to discard overexposed areas.

Limitations: The above methods [32, 54] rely on the selected thresholds, which are liable to fail in the case of relative highlight pixels. In [55], the overexposed regions are not handled efficiently. As a result, the specular regions present within the overexposed regions

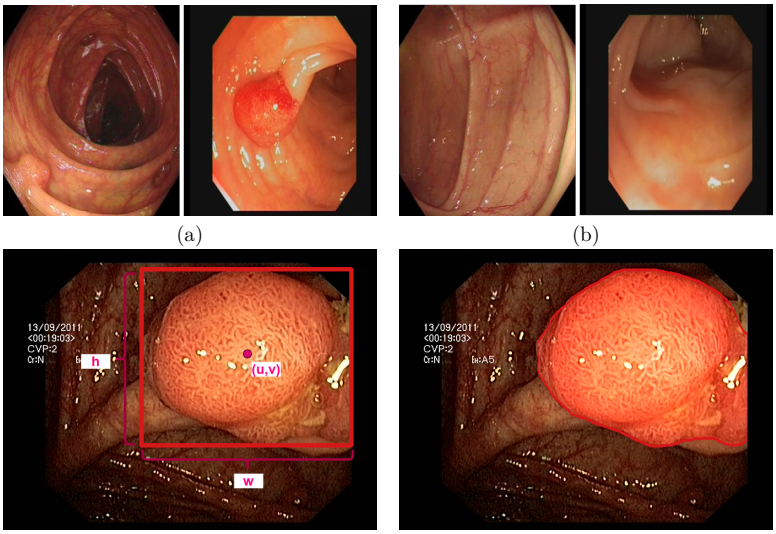


Figure 2.7: Polyp detection: (a) Polyp images vs. (b) Non-polyp images, (c) Polyp localization: denoted by bounding box, and (d) Polyp segmentation.

are not correctly detected and result in more false positives.

2.3.3 Detection and Localization

Detection and Localization are the core tasks in any computer vision based problem. Localization refers to locating the object of interest and then creating a bounding box around it. Similarly, detection represents the task of identifying objects in an image and assigning them some class. As our objective is to detect polyps, our work intends to perform these two tasks in the following sequence:

Detection: Identify a set of frames containing polyps, from a huge volume of colonoscopy videos. This process is a kind of binary classification which predicts a class of the frame as polyp or non-polyp.

Localization: Mark the exact location of the polyp in a frame using a bounding box. This process predicts a set of values $\{c, u, v, w, h\}$ where c represents the class (polyp/non-polyp), u, v, w, h denote the centre coordinates, width, and height of the bounding box around the detected polyp.

In this thesis, we use these two terms to denote *detection of polyp/non-polyp frames* from a complete colonoscopy video and *localization of polyp* in a given frame. The same is illustrated in Fig. 2.7(a)-(c).

Related Work

Polyp detection and localization have been considered independent problems for the past many years and not as a sub-part of keyframe selection. Related existing works employed handcrafted-based feature learning methodologies and relied on features like color, shape, contour, texture, etc., for the characterization of polyps [56, 57]. However, manual feature extraction needs vast domain knowledge and could be incapable of detecting polyps acquired using different imaging techniques. Therefore, deep learning based approaches have been adopted for their ability to extract hidden image features. For example, Thambawita *et al.* [58] evaluated five machine learning based models in various settings to detect polyps and 15 other classes. They also performed cross-dataset evaluations to demonstrate the challenges faced in developing a generalized model. Pacal *et al.* [59] embedded CSPNet into the YOLOv3 and YOLOv4 architecture and replaced Leaky ReLU with the SiLU activation function. They also investigated the effect of transfer learning and the inclusion of negative samples in the training data.

Considering the importance of image pre-processing, Qian *et al.* [60] detected specular reflections and adopted an image patching algorithm to remove these reflections. A tailored version of faster R-CNN is then applied on these pre-processed images to perform polyp detection. Similarly, Tian *et al.* [61] rejected frames with water-jet sprays and feces using a binary classifier and detected polyp frames as a few-shot anomaly classification problem. To benchmark various state-of-the-art methods, Jha *et al.* [62] carried out a comprehensive comparison and also proposed ColonSegNet. Though their proposed model aims at segmentation, they compared the predicted bounding boxes with other methods.

Limitations: The above detection and localization methods rarely conducted cross-dataset or patient-wise validations, which is an important requirement in real-world scenarios. Most methods have used manually selected good-quality images. Such manual refining is a time-consuming job. Other approaches fed complete video in their proposed model, which requires more computational resources.

2.3.4 Segmentation

Segmentation is a fundamental task in computer vision that partitions a given image into meaningful regions. In the colonoscopy domain, polyp segmentation refers to the delineation of the polyp region from its boundaries. In the process, labels are assigned at a pixel level, which helps demarcate polyps in a colonoscopy image. Given an image set $Q = \{I_i, O_i\}_{i=1}^{\mathtt{m}}$, where I_i represents original RGB images, and $O_i = \{o_j^{(i)}, j = 1, 2, ..., |I_i|, o_j^{(i)} \in$

$\{0, 1\}\}$ represents the corresponding masks. The pixels in the mask with $o_j^{(i)} = 1$ signify polyp regions, whereas those with zeroes belong to the background. A sample of the polyp segmentation task is shown in Fig. 2.7 (d).

Related Work

Conventional methods: In earlier years, the traditional methods adopted for polyp segmentation relied on hand-crafted features, such as texture, color, and shape. Some of these techniques are discussed in [10, 63]. However, their limited performance in handling the complex segmentation task and the inability to capture polyp heterogeneity led to the advent of deep learning approaches in this domain.

Deep Learning based methods: In recent years, several deep learning-based techniques have been developed for polyp segmentation and have reported impressive outcomes. The introduction of fully convolutional networks paved the way for many semantic segmentation models [64, 65]. Subsequently, encoder-decoder structures dominated the field with the popularity and success of U-Net [66] based architectures in biomedical image segmentation. Several studies have leveraged this concept; for example, Jha et al. [13] presented ResUNet++, which integrates residual blocks with Atrous Spatial Pyramidal Pooling, squeeze-and-excitation units, and attention blocks. Later, its extension [67] incorporated Conditional Random Field and Test-Time Augmentation for improved polyp segmentation. DoubleU-Net [14] stacked two U-Net architectures, in which the first network used a pre-trained encoder to generate features as input for the second network. Despite the excellent performance of such models, the need for real-time systems emerged due to their high resource requirements and inference time. With this focus, Jha et al. [68] proposed ColonSegNet, which produces efficient output while achieving real-time performance. Similarly, HarDNet-MSEG [15] used a low-memory traffic backbone and a cascaded partial decoder to achieve fast polyp segmentation. Focus U-Net [69] also uses an encoder-decoder architecture that embeds a dual-attention gated module and introduces hybrid focal loss.

Besides U-Net based models, several other architectures also exist that focus on boundary details and consider the camouflage property for polyps. For example, PraNet [70] comprise reverse attention modules to include boundary cues with a global feature map obtained using a parallel partial decoder. Yue et al. [71] proposed a boundary constraint network that utilizes a bilateral boundary extraction module to investigate polyp and non-polyp regions. Ta et al. [72] designed BLE-Net with an encoder deployed with a boundary learning module along with a boundary enhancement module at the decoder to consider

edge details. Polyp-PVT [73] introduced a camouflage identification module with a Pyramid Vision Transformer (PVT) encoder. This module aims to capture polyp cues concealed in low-level features. Such transformer-based approaches' excellent performance further inspired similar works in this field. ColonFormer [74] proposed a residual axial attention module and combined a hierarchical transformer with a hierarchical pyramid network for efficient polyp segmentation. HSNet [75] used PVT as an encoder and embedded cross-semantic attention module to bridge the semantic gap between low-level and high-level features. GMSRF-Net [76] developed a cross multi-scale attention module and multi-scale feature selection modules. Some other related works can be found in [77–80].

Limitations: Although the segmentation performance achieved by state-of-the-art methods is noteworthy, the issue of reproducibility and fair comparison remains. Many of the recent works have considered testing on unseen datasets. Still, the small test set size makes the validation susceptible to the kind of dataset split scheme followed, yielding biased results. Moreover, the split scheme followed in the literature is not uniform.

2.3.5 Classification

Classification is a task to identify the correct label for a given image. In the colonoscopy domain, this task can be used to recognize the pathology of a polyp that could be hyperplastic (benign) or adenomatous (potentially malignant). So, given an image set $Q = \{I_i, O_i\}_{i=1}^{m}$, where I_i represents original RGB images, O_i can be 0 or 1 depending on the pathology of the polyp present in I_i.

Related Work

Most of the prior works focused on polyp detection, and only a few targeted classification due to the unavailability of a large annotated dataset. This is why most of the methods adopted hand-crafted feature extraction techniques, as enough data was not available for training deep learning models. Only a few methods aimed at classifying polyps based on pathology (adenomatous/hyperplastic). Some methods followed a pit-pattern classification scheme to categorize normal mucosa and hyperplastic [81]. Uhl et al. [82] and Hafner et al. [83] designed fractal dimension-based schemes. The former method used two filter masks for this purpose, namely, an anisotropic Gaussian filter mask and an elliptic binary filter mask. In [83], three extensions of the local fractal dimension based approach are presented, which help in extracting shape and gradient details from the given image. Wimmer et al. [84] utilized four types of filters, three of which are directional sensitive, whereas the fourth one is

based on the maximal-minimal filter bank technique to achieve rotation invariance. Ribeiro et al. [85] first time explored different deep learning models and reported results on various model configurations (end-to-end training or transfer learning). Bora et al. [86] used the Least Square Support Vector Machine and Multi-layer Perceptron to perform classification along with different feature extraction approaches. More recently, Patel et al. [16] and Li et al. [17] released a relatively large colonoscopy dataset which is composed of different small datasets. They also provided experimental results on some baseline deep learning models. Bhamre et al. [87] explored the advantages of performing classification using NBI instead of WLI. This is achieved by translating images of WLI modality into NBI using CycleGAN.

Limitations: The literature indicates a paucity of research in the classification task, primarily attributed to the scarcity of high-quality, large-scale annotated datasets. Moreover, a research gap still persists in exploring different imaging modalities and their specific benefits in classifying polyps based on their pathology.

2.3.6 Dataset Description

Table 2.1 shows the list of the existing endoscopy datasets along with data type, size, and accessibility. It can be observed that most of the existing datasets in the literature are from colonoscopy procedures and consist of polyp still frames or videos. In this thesis, we used datasets according to the task objectives and the type of annotations available. For example, the datasets with segmentation mask and detection bounding box details are used for polyp segmentation and detection, respectively. We also designed a novel multi-class endoscopy dataset, which will be discussed in Chapter 5. The following related work discusses the existing datasets in detail to provide an overview of the datasets used in this thesis. These details also help relate to the need for our proposed dataset in Chapter 5 as most of the existing datasets are small in size and do not capture some critical anatomical landmarks or pathological findings.

Related Work

In the earlier GI detection works, the CVC-ClinicDB [10] and CVC-ColonDB [56] were widely used. **CVC-ClinicDB** is developed from 23 colonoscopy video studies acquired with white light. These videos provide 31 video sequences, each containing one polyp, which finally generates 612 images of size 576 $\times$ 768. **CVC-ColonDB** consists of 300 different images obtained from 15 random cases. Similarly, **ETIS-Larib Polyp DB** [11] is a colonoscopy dataset consisting of 196 polyp frames and their corresponding segmen-

Table 2.1: List of the existing datasets within GI endoscopy.

Dataset	Data type	Size	Accessibility
Kvasir-SEG [9]	Polyps	1,000 images†♣	Public
HyperKvasir [88]	GI findings	110,079 images & 374 videos	Public
Kvasir-Capsule [89]	GI findings◇	4,741,504 images	Public
Kvasir [90]	GI findings	8,000 images	Public
CVC-ColonDB [56]	Polyps	380 images† ‡	As per request•
ETIS-Larib Polyp DB [11]	Polyps	196 images†	Public
EDD2020 [12, 91]	GI lesions	386 images†♣	Public
CVC-ClinicDB [10]	Polyps	612 images†	Public
CVC-VideoClinicDB [19]	Polyps	11,954 images†	As per request
ASU-Mayo [92]	Polyps	18,781 images†	As per request•
KID [93]	Angiectasia, bleeding, inflammations◇	> 2500 images, 47 videos	Public•
PolypGen [94]	Polyps	1,537 images†♣ & 2,225 video sequence, 4,275 negative frame	Public
SUN Database [95]	Polyps	158,690 video frames♣	As per request
ISIT-UMR [96]	Polyps	76 colonoscopy videos	Public

†Segmentation ground truth •Not available now ‡Contour ◇Video capsule endoscopy ♣ Bounding box information

tation masks. Recently, **Kvasir-SEG** [9] dataset has been introduced that comprises of 1,000 colonoscopy images with segmentation ground truth and bounding box coordinate details. This dataset offers a diverse range of polyp frames, including multiple diminutive polyps, small-sized and regular polyps, sessile or flat polyps collected from varied cohort populations. The dataset is open-access and is one of the most commonly used datasets for polyp segmentation.

The **ASU-Mayo Clinic Colonoscopy Video (c) database** [92] is a copyrighted dataset and is considered the first largest collection of short and long video sequences. Its training set is composed of 10 positive shots with polyps inside and 10 negative shots with no polyps. The associated test set is provided with 18 different unannotated videos. **CVC-VideoClinicDB** [19] is extracted from more than 40 long and short video sequences. Its training set comprises 18 different sequences with an approximate segmentation ground truth and Paris classification for each polyp. **ISIT-UMR** [96] is a video dataset with 76 colonoscopy videos available in both NBI and WLI. It contains annotations for pathology classes, including hyperplastic (21 videos), adenomatous (40 videos), and serrated adenoma (15 videos). **SUN Colonoscopy Video Database** [95] comprises 49,136 polyp frames and 109,554 non-polyp frames. Unlike the datasets described above, this dataset includes pathological classification labels, polyp size, and shape information. It also includes bounding box coordinate details. The **PolypGen** [94] dataset is an open-access dataset that comprises 1,537 polyp images, 2,225 positive video sequences, and 4,275 negative frames. The dataset

is collected from six different centers in Europe and Africa. Altogether, the dataset provides 3,762 positive frames and 4,275 negative frames. These still images and video frames are collected from varied populations, endoscopic systems, and surveillance experts in Norway, France, the United Kingdom, Egypt, and Italy and is one of the comprehensive open-access datasets for polyp detection and segmentation.

Apart from the lower GI-related datasets, there are a few datasets that provide combined samples of upper and lower GI findings. For example, **HyperKvasir** [88] is a multi-class GI endoscopy dataset that covers 23 classes of anatomical landmarks. It contains 110,079 images out of which 10,662 are labeled and 99,417 are unlabeled images. The **EDD2020** dataset [12, 91] is a collection of five classes and 386 still images with detection and segmentation ground truth. The classes are divided into 160 non-dysplastic Barrett's, 88 suspicious precancerous lesions, 74 high-grade dysplasia, 53 cancer, and 127 polyps with overall 503 ground truth annotations. The **Kvasir-Capsule** [89] is a video capsule endoscopy dataset comprising 4,741,504 image frames extracted from 117 videos. From the total frames, 4,694,266 are unlabeled, and 47,238 frames are annotated with a bounding box for each of the 14 classes. Similarly, **KID** [93] is a capsule endoscopy dataset with 47 videos and over 2,500 images. The images are annotated for normal, vascular, inflammatory, lymphangiectasias, and polypoid lesions.

Limitations: The literature review shows that most GI-related datasets focus on a single specific finding, such as colon polyps. Some of the datasets are small in size and have ignored non-lesion frames, which are essential for developing algorithms to be integrated into clinical settings. Additionally, many of these datasets are available on request and require approval from the data providers, resulting in further delays. A few datasets like Kvasir, HyperKvasir, Kvasir-Capsule and KID provide multiple GI findings. However, Kvasir-Capsule and KID are video capsule endoscopy datasets. The Kvasir dataset has only eight classes, whereas Hyperkvasir has 23 classes.

2.3.7 Synthetic Dataset Generation

With the advent of generative artificial intelligence, several works have been proposed to generate synthetic colonoscopy images. These synthetic data mimic the characteristics of real data and fairly look similar to them. The techniques adopted so far can broadly be divided into GANs, and Diffusion Models.

Related Work

GAN based Techniques: The initial frameworks for polyp generation are based on the adversarial concept and adopt different variants of GANs. For example, Shin et al. [97] used a conditional-GAN approach to translate normal colonoscopy images to polyp images. This translation is achieved using an input-conditioned image which is a combination of an edge map and a polyp binary mask. A similar concept of converting normal frames to polyp frames is proposed in [98]. They utilized a conditional GAN architecture to produce polyps with varied characteristics by controlling the input-conditioned binary mask values. Such conditional translation is also reported by Fagereng et al. [99]. They developed a framework called PolypConnect which uses an EdgeConnect model to convert clean colon images to polyps when given an edge map and a polyp mask. Sasmal et al. [100] performed polyp generation using DCGAN and used the obtained synthetic polyps to enhance classifier performance for differentiating adenoma and hyperplastic. An identical augmentation approach is followed by Adjei et al. [101] using synthetic polyps generated using a Pix2Pix model. Unlike the traditional GAN architecture, He et al. [102] introduced an attacker in the framework to obtain false negative images. Sams and Shomee [103] utilized a StyleGAN2-ada to generate random binary masks which are combined with colon images. This integrated image is used as an input for a conditional GAN to obtain synthetic polyp images. The above methods focused on polyp generation irrespective of the imaging modalities. However, a few works used GAN-based approaches to transfer style between different imaging modalities like WLI and NBI. Golhar et al. [104] utilized the GAN inversion approach, which uses a latent representation of images to perform translation between NBI and WLI modalities. Following this technique, interpolation methods are used to change the polyp size. Similarly, Bhamre et al. [87] used CycleGAN to convert WLI images to NBI images.

Diffusion Model based Techniques: The related literature involves only a few works focused on polyp image generation. Machacek et al. [105] used a conditional diffusion probabilistic model to produce synthetic polyp images using synthetic masks. They validated the effectiveness of generated data by utilizing it for training polyp segmentation models. Pishva et al. [106] performed polyp generation using two diffusion models. The two models are fine-tuned on cropped-out polyps and clean colon images, respectively. This fine-tuning is followed by performing an inpainting using the latter model and cropped-out images. Du et al. [107] proposed an adaptive refinement semantic diffusion model which considers the polyp and background ratio to adjust the diffusion loss. They also incorporated a pre-trained segmentation model that modifies the refinement loss depending on the difference

between the predicted mask of the synthetic polyp and the actual binary mask.

Limitations: Although a few existing works have established the significance of NBI images over WLI images in polyp classification, the generation of new synthetic polyp images with different imaging modalities has not been explored in the literature. Moreover, the above-mentioned approaches followed a similar pattern of polyp generation using binary masks. The impact of text prompt based training, particularly with different imaging modalities, still remains unexplored.

2.4 Standard Metrics

Detection of polyp from a given set V of colonoscopy videos is a binary classification task, i.e., we consider two classes, polyps and non-polyps. Let tp be the number of such samples which are correctly detected as polyps and tn be the number of instances which are correctly predicted as non-polyps (background). Similarly, let fp and fn be the incorrect predictions erroneously considered polyps and background, respectively.

- *Precision:* Precision is the ratio of correctly predicted positive observations to all the positive predicted observations. It can be defined as:

$$Precision = \frac{tp}{tp + fp} \tag{2.1}$$

- *Recall/Sensitivity:* Recall/Sensitivity presents the ratio of correctly predicted positive observations to all the original observations in the actual class. It can be formulated as:

$$Recall = \frac{tp}{tp + fn} \tag{2.2}$$

- *F1-score:* F1-score integrates both recall and precision and calculates a weighted average/harmonic mean of these two metrics. It can be computed as:

$$F1\text{-}score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{2.3}$$

- *Intersection over Union (IoU)/Jaccard Index (JI):* IoU/JI quantifies the amount of overlap between the predicted and ground truth region. It can be defined as:

$$JI(Pr, Gr) = \frac{|Pr \cap Gr|}{|Pr \cup Gr|} \tag{2.4}$$

- *Average Precision (AP) and mean Average Precision (mAP):* AP can be defined as the area under a precision-recall curve which helps summarize the curve in a single value. It is a weighted sum of precisions computed over different thresholds, considering the increase in recall as the weight. mAP is defined as the average values of APs over different classes. These terms can be formulated as:

$$AP = \int_{r'=0}^{1} p'(r')\, dr' \tag{2.5}$$

$$mAP = \frac{1}{c} \sum_{i=1}^{c} AP_i \tag{2.6}$$

- *Area under Precision-Recall Curve (AUPRC):* AUPRC is the area under a precision-recall curve. It is an important metric for problems with imbalanced data where detecting positive class is the primary objective.

- *Balanced Accuracy (BA):* BA is also used with imbalanced datasets and can be calculated as the arithmetic mean of specificity and sensitivity which is given by:

$$BA = \frac{Sensitivity + Specificity}{2} \tag{2.7}$$

where $specificity = tn/(tn + fp)$.

Chapter Summary

In this chapter, we covered some of the key medical terminologies and the background concepts required to understand the subsequent chapters. These concepts are followed by literature survey pertaining to different phases of an ideal CAD system. These phases are the fundamental tasks intertwined within our proposed methodologies. The survey of each phase is accompanied by limitations, specifying the research gap in the current scenario. Also, we discussed about the different datasets and some existing synthetic dataset creation techniques, followed by their limitations. Lastly, we explained the standard metrics that are used in this thesis.

3

Keyframe Extraction for Enhanced Polyp Detection and Localization

Chapter Highlights

- We hypothesised that reviewing only some adequate number of informative frames instead of an entire colonoscopy video could yield enhanced outcomes, and significant savings in manual effort, computational resources, and time.
- We propose a multi-stage keyframe extraction framework to select good-quality, distinct polyp frames covering multi-views of polyps with a high polyp-to-frame ratio.
- We also propose *YcOLOn*, a novel multi-scale attention-based model for enhanced polyp localization performance.
- This chapter is based on the publications "*A Multi-Scale Attention Framework for Automated Polyp Localization and Keyframe Extraction From Colonoscopy Videos*" published in IEEE Transactions on Automation Science and Engineering and "*Keyframe Selection from Colonoscopy Videos to Enhance Visualization for Polyp Detection*" presented at IV 2022.

This chapter addresses the problem of extracting meaningful clinical information from colonoscopy videos, which otherwise requires an enormous amount of reviewing time and puts a considerable burden on the surgeons.

The proposed solution presents a first end-to-end automated multi-stage deep learning framework to extract an adequate number of clinically significant frames, i.e., keyframes from colonoscopy videos. The proposed framework comprises multiple stages that employ different deep learning models to select keyframes, which are high-quality, non-redundant polyp frames capturing multi-views of polyps. In one of the stages of our framework, we also propose a novel multi-scale attention-based model, YcOLOn, for polyp localization, which generates ROI and prediction scores crucial for obtaining keyframes. We further designed a GUI application to navigate through different stages. Extensive evaluation in real-world scenarios involving patient-wise and cross-dataset validations shows the efficacy of the proposed approach. The framework removes 96.3% and 94.02% frames, reduces detection processing time by 38.28% and 59.99%, and increases mAP by 2% and 5% on the SUN database and the CVC-VideoClinicDB, respectively. The source code is available at `https://github.com/Vanshali/KeyframeExtraction`.

3.1 Introduction

With about 14.2 million procedures performed in the US alone [108], colonoscopy is one of the most common tests followed to reduce CRC-related mortality and morbidity. The tremendous amount of medical data obtained from these medical examinations needs large storage volumes and enormous time for manual reviewing. This time-consuming assessment puts a considerable burden on the clinicians and introduces human errors in the diagnostic results. Some recent research proposed automated polyp detection and classification techniques [58, 62, 109–111] to reduce manual efforts. However, as many colonoscopy video frames are clinically insignificant and do not contribute to pathological interpretations, they provide distorted information to the automated systems. Hence, these systems often underperform.

One possible solution to overcome these issues is to perform video skimming that extracts the essential and good-quality content of the video streams in the form of significant and representative frames. These are called the keyframes, which preserve the static summaries of the key-events corresponding to a patient's colonoscopy procedure. In this way, an adequate number of good-quality frames can be obtained. The automated systems can then

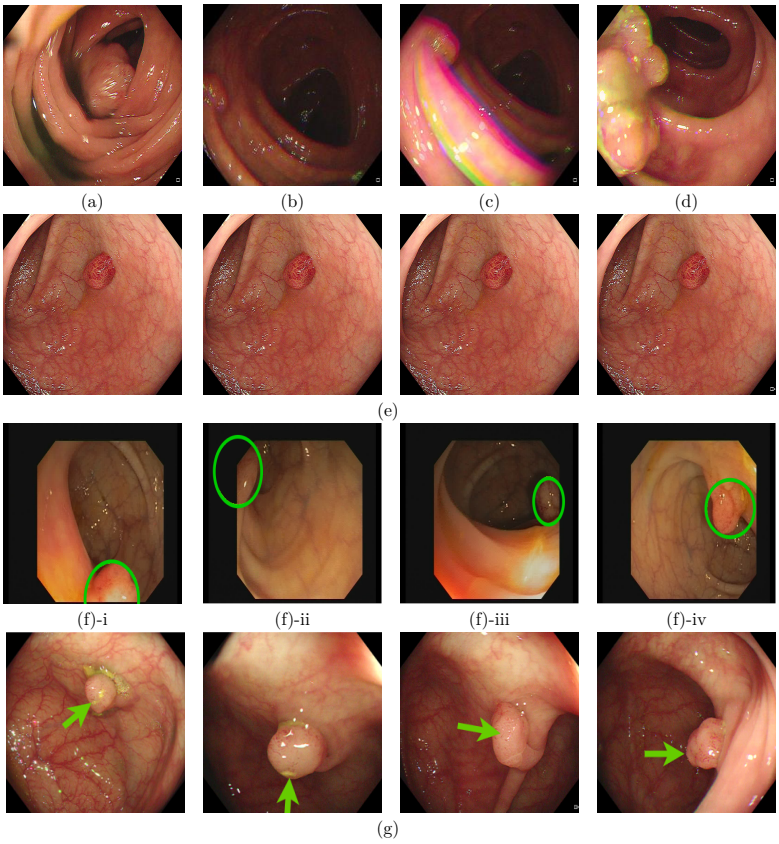


Figure 3.1: The figure illustrates example scenarios of frames captured in the colonoscopy procedure. The first row displays low-quality frames: (a) motion blur, (b) low illumination, (c) ghost colors, and (d) waste materials. The second row (e) displays redundant consecutive frames from a video sequence. The third row (f) displays images of the same polyp, where (f)-i and (f)-ii show cropped view, (f)-iii shows a distant view, and (f)-iv shows a clear view. The fourth row (g) displays frames with multiple views of a polyp from different directions.

efficiently process the extracted keyframes and easily store them for clinical assessments. The keyframes can save doctors a lot of reviewing time and allow them to go through many case studies and surgeries quickly. Some keyframes can also be attached to a patient's medical record file for future reference and retrospective treatments.

An effective keyframe selection approach must consider various attributes of the recorded colonoscopy videos [43]. Firstly, about 25% of the acquired video frames are low-quality [35]. These frames can be blurry or have noise such as interlacing, ghost colors, low illumination, fecal deposition, and overexposed regions (Fig. 3.1(a)-3.1(d)). Secondly, most consecutive frames are redundant [112] and do not add to the clinical findings (Fig. 3.1(e)). Thirdly, not all polyp frames provide a satisfactory Polyp-to-frame ratio (PI), and some present cropped or distant polyp views (Fig. 3.1(f)). A colonoscopy keyframe selection procedure should consider these attributes to provide clinicians with essential data from a

large set of recordings. In addition to the quality frames, a clinician looks for multiple views of the detected polyps for efficient analysis and precise resection (Fig. 3.1(g)). Therefore, keyframes must contain diverse views of each polyp.

Most polyp detection, localization, and classification methods in the literature [56, 62] either use the raw form of videos or perform manual removal of noisy frames, which is a very time-consuming task. Only a few methods [32, 33] have focused on keyframe selection in the colonoscopy domain, and they have the following drawbacks:

1. They do not consider all essential video attributes.

2. They mainly depend on the hand-crafted features, such as threshold, entropy-based frame quality assessment, sub-shot formation, and clustering-based representative frames selection.

Thus, the existing methods are not robust enough to represent high-level semantic information and overlook important colonoscopy video attributes.

In this chapter, we propose an automated deep learning based keyframe selection approach that extracts keyframes from colonoscopy videos, overcoming the above-mentioned drawbacks. The proposed approach consists of five stages, wherein each stage introduces different methods to ensure that the keyframes obtained are clinically significant. **Stage-I** eliminates the uninformative frames containing motion blur, ghost colors, fecal deposition, and low illumination using a deep learning based model. **Stage-II** performs polyp detection to retain only the polyp frames that are important for pathological interpretations. In **Stage-III**, we group redundant and highly correlated frames, enforcing diversity among the polyp frames. In **Stage-IV**, we localize the polyp region using our proposed multi-scale attention-based localization model, YcOLOn. Lastly, we use the weighted score method to

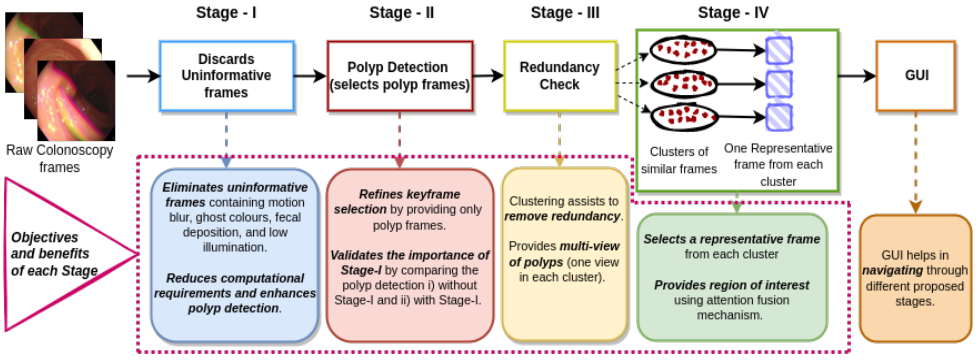


Figure 3.2: Flowchart depicting the role of different stages in the proposed work.

select one polyp frame from each group. The framework is evaluated on two publicly available datasets with patient-wise/case-wise or cross-dataset settings. The results illustrate the enhanced polyp detection and localization performance with a significant reduction in the total number of frames that need to be reviewed. The stage-wise outcomes could be visualized using the GUI application that we developed for easy navigation. Figure 3.2 illustrates the complete overview of the proposed method.

3.1.1 Contributions

The contributions of this chapter are discussed below:

- **First deep learning based framework for keyframe selection in colonoscopy domain that considers all video attributes:** To the best of our knowledge, this is the first deep learning based framework that aims to extract keyframes from colonoscopy videos while considering all the mentioned video attributes. Also, our method selects the best frame from correlated frames and enforces diversity among the keyframes. Such selection provides each polyp view from multiple directions and helps surgeons determine the exact information for polyp resection.

- **Cross-dataset validations and case-wise analysis:** We performed all the experiments either using patient-wise/case-wise data or in cross-dataset settings. This makes the proposed framework generalized and suitable for different patients' data acquired in the same hospital (with similar imaging modalities but different patient-specific characteristics) and also across different hospitals (with different imaging modalities). We performed the exhaustive case-wise analysis with each video sequence results interpreted using separate boxplots.

- **Enhanced polyp detection performance with less processing time**: We achieved enhanced polyp detection rates along with reduced computational time by adopting an approach to discard uninformative frames. (Stage-II)

- **Proposed a novel attention-based model YcOLOn for enhanced polyp localization performance:** We introduced an attention fusion mechanism that provides a more precise localization of polyps. The obtained predicted scores assist in selecting the representative frames from each cluster. (Stage-IV)

- **Annotated colonoscopy images based on the quality of the frames:** We annotated the polyp frames of SUN database [95] that could facilitate the researchers

with training labels for future research.

3.2 Proposed Method

3.2.1 Stage-I: Quality Assessment

Many artifacts, including noise, ghost colors, motion blur, fecal depositions, and low illumination, degrade the quality of colonoscopy frames leading to an incorrect diagnosis. Such frames often distract the clinician's focus and also make automated diagnostic systems underperform. We consider these frames uninformative as they do not hold clinically important information. Such frames can not be a part of keyframes and could unnecessarily impact the outcomes of the subsequent stages of the proposed pipeline. To get rid of them, we trained the existing architecture of the DenseNet-201 [113] model as a binary classifier using manually annotated informative/uninformative frames. DenseNet-201 is a Convolutional Neural Network (CNN) with 201 layers, in which each layer is connected to every other layer for richer patterns and diversified features. This dense connectivity can be illustrated as:

$$o_l = w_l * (q_0|q_1|...|q_l) \tag{3.1}$$

where o_l and w_l denote the current feature map and the convolution weight, respectively. q_i represents each of the previous layers' feature maps, and '|' indicates the concatenation operation. As an end result, we obtain:

$$V_{Stage\text{-}I} = V/V_{insignificant} = \{f \in V : f \notin V_{insignificant}\} \tag{3.2}$$

where $V_{insignificant}$ represents a set of insignificant frames.

As the output of this stage will affect the outcomes of subsequent stages, performance evaluation is required at this point. Due to the unavailability of ground truth information in one of our test datasets, two metrics, namely, the Fast Fourier Transform (FFT) [114] and the Variance of Laplacian (VoL) [115], are adopted to validate the effectiveness of this quality assessment step. These metrics quantify the blurriness content of the given image. An image can be represented as real and imaginary components using FFT. The lower the amount of high frequencies contained in these components, the higher the blurriness content. Similarly, the second metric considers an image blurry based on the rapid intensity changes. A high variance represents an image with sharp and clear content, whereas a low variance indicates an indistinct image. Suppose fft_i and vol_i are the mean of magnitude

spectrum of FFT and focus measure of VoL, respectively, obtained on the set V. Similarly, fft_j and vol_j are associated with computations performed on $V_{Stage\text{-}I}$, i.e., on the frames obtained after applying DenseNet-201. To obtain better quality frames in $V_{Stage\text{-}I}$ compared to the set of raw frames V, our objective is to achieve $fft_j > fft_i$ and $vol_j > vol_i$.

3.2.2 Stage-II: Polyp Detection

The main pathological interpretations are performed on polyp frames to diagnose the correct medical state. Therefore, after removing insignificant frames, we addressed the detection of polyp frames, V_{P_frames}, using a binary classification approach. The aim of this stage is to obtain:

$$\begin{aligned} V_{Stage\text{-}II} &= V_{Stage\text{-}I} / V_{NP\_frames} \\ &= \{f \in V_{Stage\text{-}I} : f \notin V_{NP\_frames}\} \end{aligned} \tag{3.3}$$

where V_{NP_frames} represents a set of non-polyp frames; P and NP stand for polyp and non-polyp, respectively. Let $D' = \{p_t, x_t, y_t\}_{t=1}^{|D'|}$ be the dataset, where p_t is the patient/case identification number, x_t denotes the RGB colonoscopy frame and $y_t \in \{0, 1\}$ is the corresponding frame label (either polyp or non-polyp). p_t is mutually exclusive in the training D'_{train}, validation D'_{val}, and test D'_{test} sets. Our motive is to find only that subset D_P of test data samples which contain polyps, i.e., $D_P = V_{P\_frames} \cap D'_{test}$. This process provides two-fold benefits: firstly, it refines the keyframe selection process by eliminating non-polyp frames, and secondly, it validates the effectiveness of removing uninformative frames in Stage-I. Different components of these two stages are shown in Fig. 3.3.

The input to the detection model is the output frames of the previous stage. This ensures the model gets good-quality and significant data samples to avoid unnecessary processing requirements. The model used here is the same CNN architecture that was

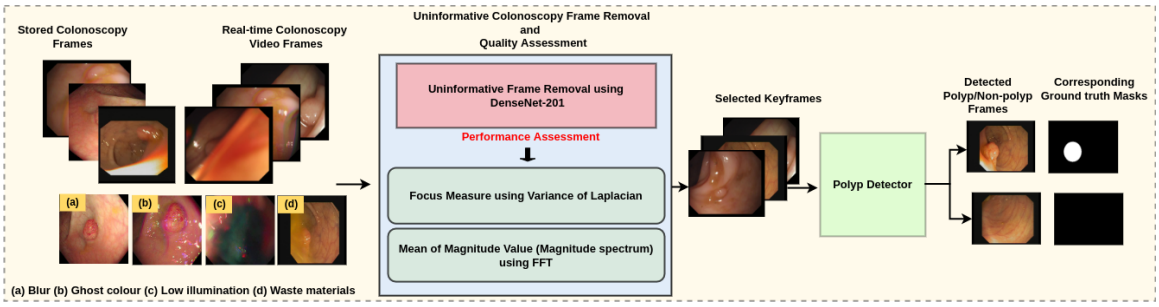


Figure 3.3: Different components of Stage-I and Stage-II.

used in Stage-I, i.e., the DenseNet-201. This network is chosen because of its remarkable performance in the detection problems [116, 117]. Moreover, during Stage-I and II, the risk of missing a critical frame is less, and unwanted frames to be processed are large. To keep it simple and less time-consuming, a basic existing CNN model is adopted, which is efficient enough at the same time for the defined task. The model is trained using two loss functions: Binary Cross Entropy (BCE) and focal loss, one at a time. Experiments with the same settings are done using ResNet-152 [118] to evidently show that the enhancement in polyp detection due to insignificant frames removal is not network or loss-dependent.

3.2.3 Stage-III: Redundancy Removal

The consecutive frames in a colonoscopy video share significantly correlated content. Compared to the general videos, the similarities encountered in the adjacent colonoscopy frames are more due to the slow motion of the cameras during the colonoscopy procedure. This raises the amount of redundancy in the video sequences. As a majority of the non-polyp frames are already removed during the previous stage, redundancy most likely exists in the small video shots containing polyps. Therefore, Stage-III addresses the issue of redundant polyp frames by training the existing architecture of the Siamese network. It is a CNN architecture that comprises two identical sub-networks, both possessing the same configuration, parameters, and weights. All parameter updates are mirrored across both the sub-networks. This ensures that two almost similar colonoscopy images are not mapped to extremely different feature spaces.

Consider colonoscopy frames f_i and f_j and let $y \in \{0, 1\}$ be the label that indicates the similarity/dissimilarity between the two frames. The Siamese network uses a metric learning approach to find the relative distance between the given input frames. In this context, a dissimilarity score is generated using the contrastive loss, which can be computed as:

$$(1 - y)\frac{1}{2}J^2 + y\frac{1}{2}\{max(0, z - J)\}^2 \tag{3.4}$$

where J is the Euclidean distance that can be defined as $(E'(f_i) - E'(f_j))^{\frac{1}{2}}$, where E is the output of the network for one image. z is a margin value beyond which dissimilar pairs will not contribute to the loss.

On the basis of the obtained dissimilarity scores, grouping is performed. Adjacent frames f_i and f_j with dissimilarity score $S < T$ are placed together under the same cluster $\mathfrak{C}$, where T is a threshold that could be set according to the required number of distinct

frames. This groups redundant frames in one cluster, providing us with:

$$V_{Stage\text{-}III} = V_{Stage\text{-}II_{\mathfrak{C}_1}} \cup V_{Stage\text{-}II_{\mathfrak{C}_2}} \cup ... \cup V_{Stage\text{-}II_{\mathfrak{C}_k}} \tag{3.5}$$

where $\mathfrak{C}_1, \mathfrak{C}_2, \mathfrak{C}_3 ... \mathfrak{C}_k$ are different clusters. This approach provides two benefits: a) we get distinct clusters with correlated frames in the same cluster, which signifies that one most significant frame from each cluster would suffice our criteria of keyframe selection, and b) polyp shots from different angles will be distributed in different clusters, and selecting one representative frame per cluster would provide varied views of a polyp.

3.2.4 Stage-IV: Polyp Localization

The selection of representative frames from each cluster considerably depends on the extent of polyp characterizing features contained within the frame. These features include the texture, color, and shape of the polyps, which make them relatively distinguishable from other endoluminal objects. One approach to quantify the presence of such features is to compute the *Confidence Score (CS)* of polyp detection that generates the probability of the polyp being detected in the frame. The higher the polyp detection CS in a frame, the greater the probability that the polyp characterizing features exist in that particular frame. Therefore, in Stage-IV, we perform polyp detection with localization, providing us with both ROIs and the corresponding CSs.

This chapter treats the above-mentioned task as a regression problem. Consider a dataset $D' = \{p_t, x_t, bb_t, y_t\}_{t=1}^{|D'|}$, where p_t, x_t, and y_t are same as defined in Section 3.2.2. $bb_t \in \mathbb{R}^4$ represents the 2-D coordinates of the bounding box containing polyp. A subset of it is used to train a detector that predicts the bounding box coordinates and corresponding CSs. Although many types of detectors are currently available, including one-stage, two-stage, anchor-based, and anchor-free detectors, a general base structure is followed in all the types. It comprises a backbone for feature extraction, a neck to make rational use of the extracted feature maps, and a head to detect the location and class of objects. Considering the one-stage detection capability with remarkable performance, we selected YOLOv5 [119] as our baseline. Its architecture consists of CSPDarkNet with an Spatial Pyramid Pooling (SPP) layer as the backbone, Path Aggregation Network (PANet) as the neck and YOLO detection head. Here, PANet is used as a link between the backbone and the head to boost information flow. It incorporates bottom-up paths and performs concatenation, which helps propagate low-level features and perform bridging of features.

However, adopting direct feature concatenation propagates features with fixed weights, ignoring the variance of feature map contents.

Therefore, we proposed a novel attention-based model, YcOLOn, which is a modified version of YOLOv5. The architectural details of YcOLOn are shown in Fig. 3.4. Unlike YOLOv5, the YcOLOn embeds four Attention Feature Fusion (AFF) [120] modules into the neck component. These modules replace the traditional method of feature concatenation to improve feature fusion and ultimately enhance polyp localization.

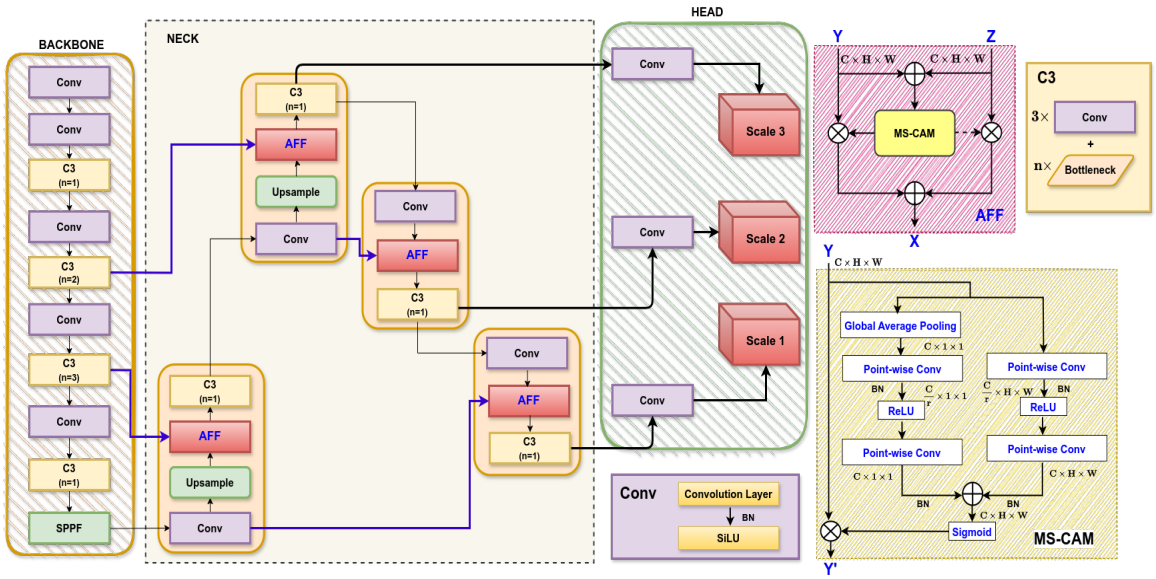


Figure 3.4: Different components of the proposed model. AFF is the attention feature fusion module, and MS-CAM is the multi-scale channel attention component of the AFF.

The different feature fusion strategies generally followed in deep learning include addition (ResNet, FPN), concatenation (U-Net, YOLOv5), attention-based refinement (SENet), and modulation (GAU). The first two strategies are not context-aware, and the other two are partially context-aware. Moreover, the channel attention module adopted in SENet emphasizes large objects while the signals pertaining to small objects vanish. Such global feature context-based components could not efficiently handle small objects like polyps. The AFF module considers multi-scale feature contexts to overcome the issues related to feature integration at different scales and vanishing signals of small objects. It incorporates a Multi-scale Channel Attention (MS-CAM) component, which combines the local channel context $L(Y)$ with the global channel context $G(Y)$, as expressed below:

$$Y' = Y \otimes MS(Y) = Y \otimes \sigma(L(Y) \oplus G(Y)) \tag{3.6}$$

where $Y', MS(Y) \in \mathbb{R}^{C\times H\times W}$ indicate the refined feature map and attention-based weights, respectively, generated by MS-CAM. $\otimes$, $\oplus$ represent the element-wise multiplication and the broadcasting addition, respectively, and σ denotes the sigmoid function. $L \in \mathbb{R}^{C\times H\times W}$ can be computed as:

$$L(Y) = BN(PwConv_2(ReLU(BN(PwConv_1(y))))) \tag{3.7}$$

where BN is the batch normalization, $PwConv_1$ and $PwConv_2$ signify the point-wise convolutions that use the channel reduction ratio r. Similarly, G can be expressed as:

$$G(Y) = BN(PwConv_2(ReLU(BN(PwConv_1(g(Y)))))) \tag{3.8}$$

where $g(Y) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} Y_{[:,i,j]}$ is the global average pooling. The same shape of L(A) as that of the input helps preserve the fine details in the low-level features. Let Y be the low-level feature map, and Z be the high-level semantic feature map, and $Y, Z \in \mathbb{R}^{C\times H\times W}$, the AFF can be computed as:

$$X = MS(Y + B) \otimes Y + (1 - MS(Y + Z)) \otimes Z \tag{3.9}$$

where $+$ is the element-wise summation and $X \in \mathbb{R}^{C\times H\times W}$ is the fused feature. The fused features get processed by the C3 blocks and are provided to the prediction heads for the final localization of polyps.

3.2.5 Final Representative Frame Selection

The final selection of representative frames is performed by determining the most significant frame from each cluster based on three parameters: a) High CS of ROI detection, b) Large PI, and c) More centrally located ROI, i.e., less distance from image's center (CD). Their corresponding values are obtained as the output of our proposed localization model as $bb_i \in \mathbb{R}^6 = \{c_i, u_i, v_i, w_i, h_i, cs_i\}$, where c_i is the class, (u_i, v_i) is the 2-D coordinate of the central pixel of the ROI, w_i, h_i are the width and height of the ROI, respectively and cs_i is the confidence score. Here, CS defines the polyp characterizing features, making it the most crucial parameter among all. However, besides the CS, the polyp area covered by the predicted bb_i and its location is also considerably important. The same polyp with greater w_i and h_i values is probable to convey more clinical information than its redundant counterpart with lower w_i and h_i values. Similarly, polyps closer to the central pixel of

the colonoscopy image indicate a relatively more visible polyp area without any cropped regions. Hence, we considered PI and CD, where $PI \propto w_i \times h_i$ ensures larger ROI and $CD = ((u_i - u_c)^2 + (v_i - v_c)^2)^{\frac{1}{2}}$ confirms a more centrally located ROI by computing distance of (u_i, v_i) from the image's center (u_c, v_c). Finally, a score is generated by jointly considering these parameters using weights α, β, and γ assigned to them, and can be given by:

$$FinalScore = \alpha \times CS + \beta \times PI + \gamma \times (1 - CD) \tag{3.10}$$

where α, β, and γ are set to 0.5, 0.3, and 0.2, respectively, providing more significance to CS followed by PI and CD.

3.3 Experimental Results

3.3.1 Dataset Details and Training Settings

The proposed work uses the publicly available SUN database [95, 121] and CVC-VideoClinicDB dataset [122, 123]. We used the SUN database for training purposes in all the experiments. It consists of annotated video frames, out of which 109, 554 are non-polyp frames from 13 cases, and the remaining 49, 136 polyp frames pertain to 100 different lesions. We split the dataset randomly into a ratio of approximately 80, 10, and 10 for training, validation, and test set, respectively. These are case-wise splits, i.e., no two splits hold data of the same case. Also, experts manually annotated polyp frames as informative or uninformative for Stage-I processing. Thus, the polyp frames have additional annotations where around 31% of the frames are uninformative. Though the proportion of each type of artifact is unknown, random sampling and case-wise division avoid biased distribution across splits. For cross-dataset performance validation, we used the CVC-VideoClinicDB, which consists of 18 video sequences. As the CVC-VideoClinicDB test set does not contain ground truths, we have used only training and validation sub-parts from Stage-II onwards.

The proposed method is implemented using PyTorch using the Titan Xp GPU. Stage I and Stage II involve 20 epochs of training, a 0.001 initial learning rate, and an Adam optimizer. In Stage III, the model is trained for 100 epochs using an Adam optimizer with an initial learning rate of 0.0005. Our proposed model in Stage-IV is trained for 80 epochs using an Stochastic Gradient Descent (SGD) optimizer with an initial learning rate set to 0.05.

3.3.2 Performance Evaluation

Stage I

The models in this stage are trained using manually annotated informative/uninformative frames of the SUN database. The validation is performed using FFT and VoL. The related results are shown in Table 3.1. The higher values of the two metrics infer less blurriness content and sharper and clearer images. It can be seen that after processing the raw frames through Stage-I models, the overall dataset quality has improved in most of the cases. This improvement is more when tested on the SUN database than the CVC-VideoClinicDB. One reason could be the similar data distribution of the training samples and the former's test set. It could also be inferred that the SUN database consists of more uninformative frames. Such low-quality frame count might be due to the large number of non-polyp frames in this database, which are expected to carry more artifacts due to unfocused camera movements. The importance of this stage is further validated in the subsequent stages, where improvement in polyp detection and localization is reported in terms of performance and computational requirements.

Table 3.1: Comparative analysis of frames quality after elimination of uninformative frames in Stage-I. + and - signs represent the value by which the frame quality improved or degraded, respectively, using Stage-I as compared to the raw frames.

Method	**Model**	**Loss**	**CVC-VideoClinicDB train_valid**	**CVC-VideoClinicDB test**	**SUN database**
			Stage-I output frames ($\pm$ δ)		
Variance of Laplacian (Focus measure)	DenseNet-201	BCE	138.26 (+1.81)	162.63 (+6.17)	480.91 (+96.79)
		Focal	**140.49** (+4.04)	165.97 (+9.51)	**487.73** (+103.61)
	ResNet-152	BCE	135.87 (-0.58)	157.11 (+0.65)	480.32 (+96.2)
		Focal	138.63 (+2.18)	**167.11** (+10.65)	479.58 (+95.46)
FFT (Mean of magnitude spectrum)	DenseNet-201	BCE	-16.59 (+0.7)	**-18.81** (+1.2)	34.98 (+19.05)
		Focal	**-14.32** (+2.97)	-18.97 (+1.04)	35.56 (+19.63)
	ResNet-152	BCE	-17.30 (-0.01)	-20.00 (+0.01)	**35.69** (+19.76)
		Focal	-16.10 (+1.19)	-19.98 (+0.03)	34.66 (+18.73)

It is noteworthy that apart from the overall performance, patient-wise results are also crucial. Similar to the empirical rule by Ma et al. [43], we followed a criterion to retain at least 20% frames of each video sequence during Stage-I testing. The purpose of setting such criteria is to avoid the lack of keyframes for correct clinical decision-making. This would be of great help in case the collected video is extremely unacceptable in terms of frame quality due to unsatisfactory patient preparation or unexpected events during the colonoscopy procedure. Considering such scenarios, this condition is set only for this particular stage. The criterion can be fulfilled by tuning the CS of the model during testing and can easily be done by clinicians. We have not laid any upper limit on the number of frames in Stage-I because of the elimination process associated with the subsequent stages. To test our

approach without any tuning, we assessed the overall performance with default settings, irrespective of the case-specific data. However, while analyzing the case-wise performance, we inspected the number of frames in Stage-I and tweaked the CS so that 20% of the patient's frames were retained.

Table 3.2: Impact on polyp frame detection results after discarding uninformative frames.

	Method	Loss	SUN database				CVC-VideoClinicDBtrain_valid			
			Precision (%)	Recall (%)	F1-score (%)	Time (sec)	Precision (%)	Recall (%)	F1-score (%)	Time (sec)
Before discarding uninformative frames (Raw frames)	ResNet-152	BCE	78.17±1.45	78.94±1.88	78.53±0.20	985	85.11±0.21	95.37±1.68	**89.94±0.63**	2076
		Focal	77.23±6.12	74.08±7.42	75.32±0.94	982	88.09±1.14	82.28±5.08	85.02±2.18	2078
	DenseNet-201	BCE	63.81±16.6	83.80±11.6	71.03±6.43	389	87.03±0.74	88.59±2.89	87.78±1.05	600
		Focal	82.30±0.16	73.94±0.03	77.89±0.06	734	**88.26±2.95**	89.07±10.43	88.40±3.69	1877
After discarding uninformative frames (Stage-I output frames)	ResNet-152	BCE	**84.28±2.34**	75.59±2.79	79.65±0.51	374	82.63±0.03	**96.96±1.25**	89.22±0.55	821
		Focal	80.14±6.16	71.98±9.62	75.42±2.59	388	84.89±1.43	87.95±0.29	86.39±0.62	820
	DenseNet-201	BCE	70.51±15.0	**83.93±16.8**	75.01±1.85	**156**	84.44±1.41	93.40±2.21	88.68±0.22	**292**
		Focal	83.51±0.98	76.38±1.07	**79.78±0.13**	453	85.28±2.40	92.86±6.67	88.80±1.76	751

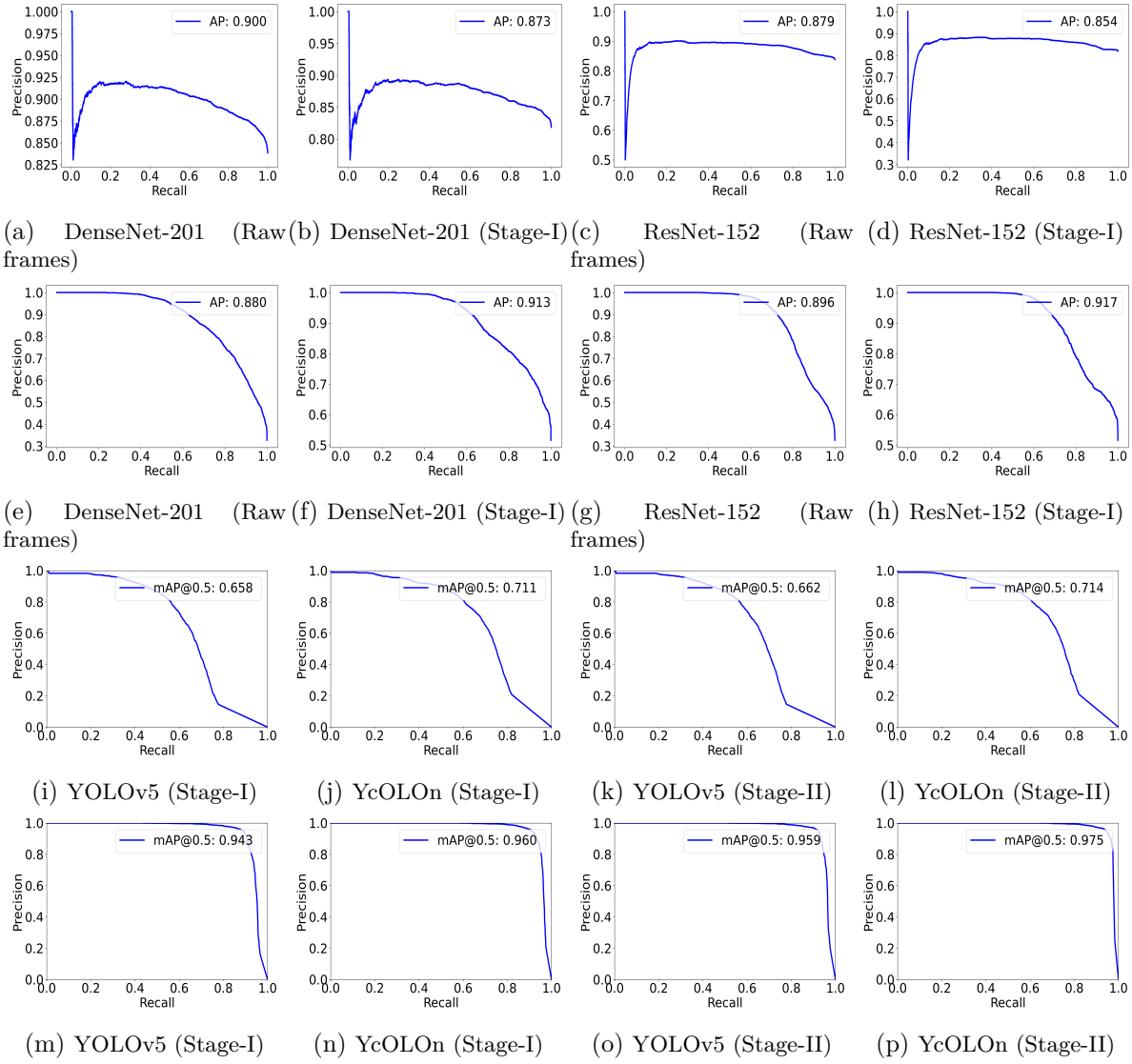
Table 3.3: Comparative analysis of our polyp detection performance (Sensitivity %) with state-of-the-art methods.

Method	SUN database							CVC Dataset
	Case-90	Case-73	Case-61	Case-59	Case-51	Case-13	Case-11	CVC-VideoClinicDB
Misawa et al. [95]	97.7	93.5	92.1	96.9	91.5	92.2	95.6	-
Thambawita et al. [58]	-	-	-	-	-	-	-	94.86
Ours	100	91.27	99.13	99.56	99.70	100	57.69	96.96

Stage-II

We carried out two types of polyp detection experiments: a) using raw frames and b) using processed frames obtained from Stage-I. Table 3.2 and AUPRC in Fig. 3.5 (a)-3.5 (h) show the comparison between these two different scenarios. Results present comparable or improved detection rates using our approach in terms of precision, recall, and F1-score, along with less processing time involved. The substantial reduction in the computational time is due to a 38.16% and 61.62% decrease in the number of testing frames of the CVC-VideoClinicDB and the SUN database, respectively. Fig. 3.6 provides a patient-wise analysis of 18 video sequences in CVC-VideoClinicDBtrain_val. It can be observed that for each patient, the number of frames and processing time drastically decreased. At the same time, sensitivity is increased or is comparable to that obtained using raw frames. The two models, DenseNet-201 and ResNet-152, demonstrate a similar trend of performance improvement. However, we utilized the polyp frames obtained from the DenseNet-201 due to its lower processing time and satisfactory detection results.

To illustrate the performance of Stage-II independently, we compared the results with



(a) DenseNet-201 (Raw frames) (b) DenseNet-201 (Stage-I) (c) ResNet-152 (Raw frames) (d) ResNet-152 (Stage-I)

(e) DenseNet-201 (Raw frames) (f) DenseNet-201 (Stage-I) (g) ResNet-152 (Raw frames) (h) ResNet-152 (Stage-I)

(i) YOLOv5 (Stage-I) (j) YcOLOn (Stage-I) (k) YOLOv5 (Stage-II) (l) YcOLOn (Stage-II)

(m) YOLOv5 (Stage-I) (n) YcOLOn (Stage-I) (o) YOLOv5 (Stage-II) (p) YcOLOn (Stage-II)

Figure 3.5: The AUPRCs are associated with Stage-II (a)-(h) and Stage-IV (i)-(p) outcomes. (a)-(d) and (i)-(l) present results using CVC-VideoClinicDB. Similarly, (e)-(h) and (m)-(p) show curves for the SUN database outcomes.

the state-of-the-art polyp detection methods. Very few works [58, 95] have used the same dataset as ours. Though not exactly comparable, we have tried to conduct a fair comparison as shown in Table 3.3. Misawa et al. [95] have reported case-wise per-frame sensitivity on the SUN database. As we used a subset of this dataset for testing purposes, we have shown results on a common set of patients' data. Thambawita et al. [58] experimented in multiple settings, so we compared with their best outcome. The comparative analysis shows that our approach outperformed other methods in all the cases except for cases 11 and 73. The reason could be these patients' very small-sized polyps (3mm) [95]. Nevertheless, the overall improvement validates that Stage-I processing reduces the computational requirements and

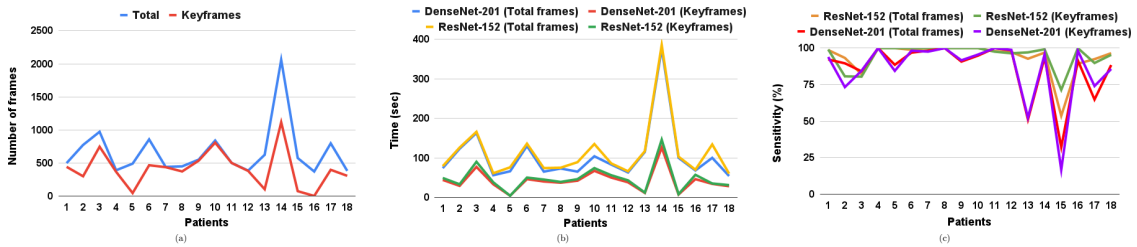


Figure 3.6: Graphs representing all patient-wise changes observed after keyframe selection. (a) shows the decrease in the number of images that need to be tested for polyp detection, (b) presents the reduction in both the models' testing time, and (c) provides an overview of enhanced overall polyp detection rate in terms of sensitivity of the two models.

also enhances the polyp detection rate and the dataset quality.

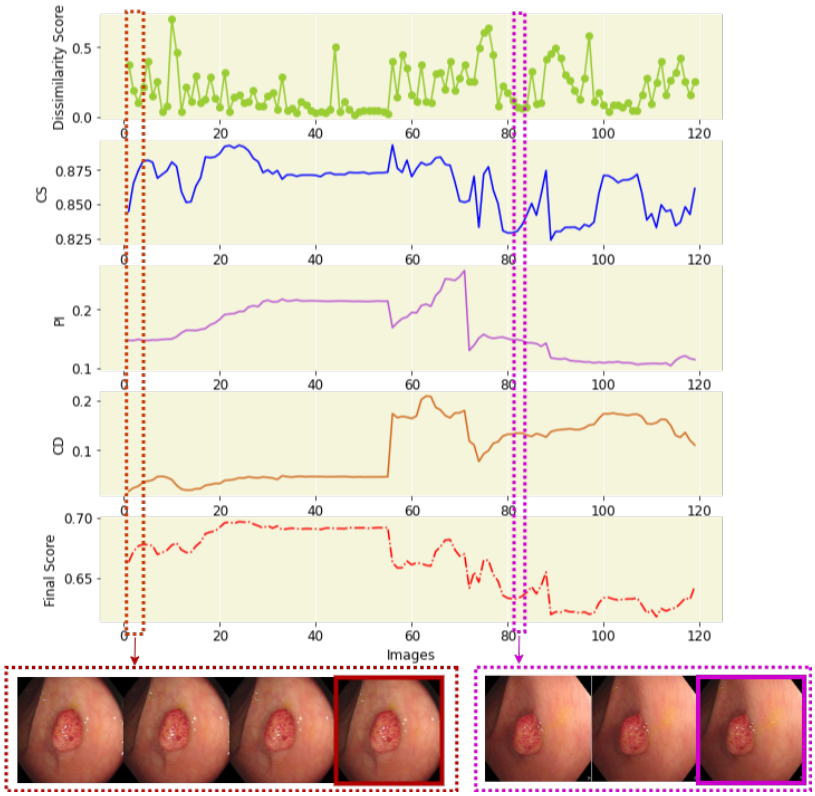


Figure 3.7: The graph plots the dissimilarity scores, CS, PI, CD, and final scores of a video shot segmented from a patient's video sequence. The two dotted boxes over the sub-plots represent two independent clusters. The keyframe selected from each cluster is highlighted.

Stage-III

This stage computes the dissimilarity scores using pairs of two immediate adjacent frames. An example is shown in the first sub-plot of Fig. 3.7. The graph depicts the dissimilarity scores pertaining to the video shot of a patient in the SUN test set. The lower scores

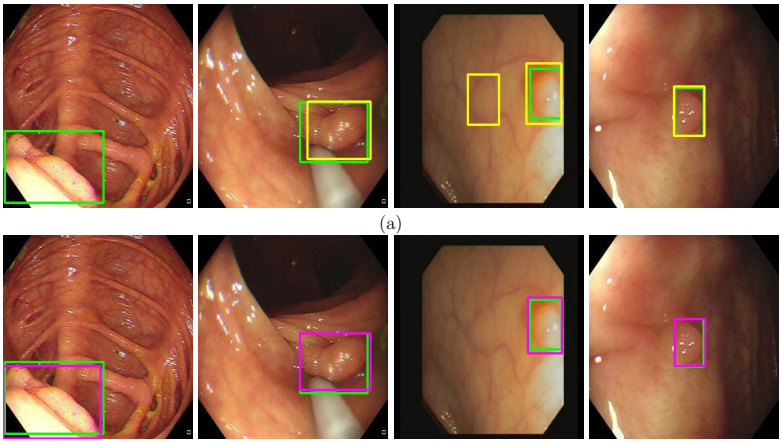


Figure 3.8: Sample images illustrating the comparative analysis of the localization performance: (a) YOLOv5, and (b) YcOLOn. The green, yellow, and pink color bounding boxes denote the ground truth, YOLOv5, and YcOLOn predictions, respectively.

Table 3.4: Comparative analysis of polyp localization results.

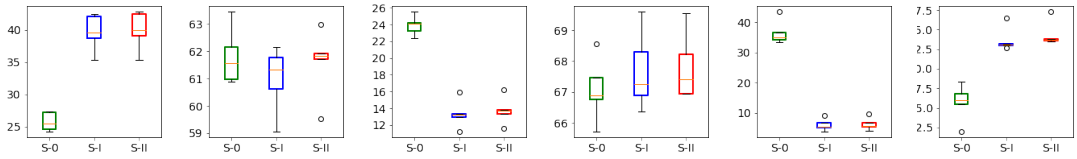
	Method	SUN database					CVC-VideoClinicDBtrain_valid				
		Precision (%)	Recall (%)	F1-score (%)	mAP @0.5	mAP @0.5:0.95	Precision (%)	Recall (%)	F1-score (%)	mAP @0.5	mAP @0.5:0.95
Before discarding non-polyp frames	YOLOv5	93.08±1.80	89.57±0.52	91.28±0.59	94.51±0.24	58.43±1.41	74.88±4.75	58.35±2.12	65.51±1.84	63.26±3.9	31.09±1.64
	YcOLOn	94.00±1.45	91.33±0.57	92.63±0.45	95.55±0.61	60.69±0.29	**76.47±3.33**	64.34±1.06	69.85±0.74	71.22±0.18	36.57±0.27
After discarding non-polyp frames	YOLOv5	94.32±1.40	92.42±0.85	93.35±0.35	96.39±0.41	60.82±1.36	75.37±4.31	58.64±2.82	65.87±1.69	63.74±3.82	31.24±1.64
	YcOLOn	**95.22±1.17**	**93.32±1.19**	**94.25±0.60**	**96.81±0.63**	**62.88±0.14**	74.88±4.26	**65.99±2.01**	**70.07±0.67**	**71.41±0.10**	**36.6±0.37**

indicate a higher correlation of the current frame with the previous one, and an abrupt increase in the scores signifies a non-redundant frame. This behavior of the video sequence scores helps form clusters of redundant frames. As this correlation is patient-specific and setting the same threshold for all the video sequences will not provide relevant clusters, we followed a case-wise approach to set cluster boundaries. For each case, an upper quartile of dissimilarity scores is calculated. Any score above this value indicates a non-redundant frame and the starting point of a new cluster. This process ensures that the number of clusters obtained is about 25% of the total polyp frames. Thus, selecting one representative frame from each cluster puts an upper limit on the number of keyframes, i.e., 25% of a patient's detected polyp frames.

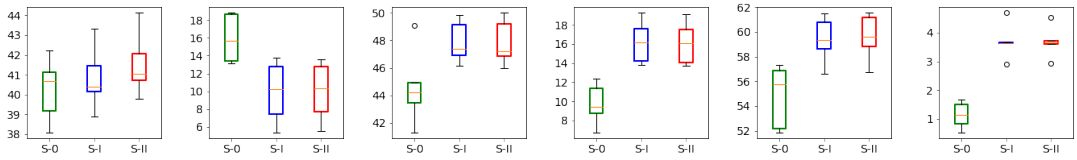
Stage-IV

The experiments in this stage are conducted using our proposed model described in Section 3.2.4, and a comparison is drawn with the baseline model, YOLOv5.

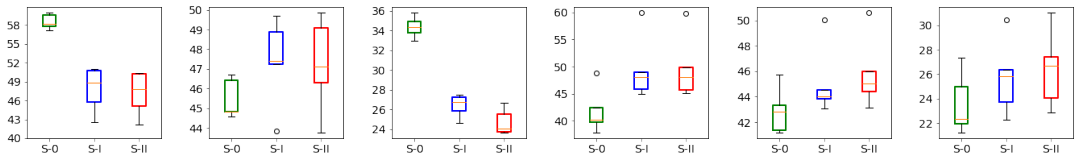
Comparison with the baseline: To investigate the effect of the AFF module, we



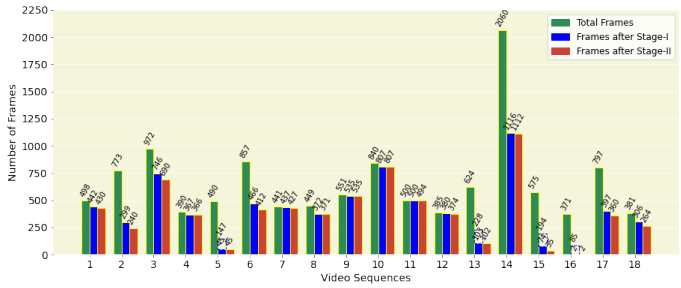
(a) Sequence-1 (b) Sequence-2 (c) Sequence-3 (d) Sequence-4 (e) Sequence-5 (f) Sequence-6



(g) Sequence-7 (h) Sequence-8 (i) Sequence-9 (j) Sequence-10 (k) Sequence-11 (l) Sequence-12



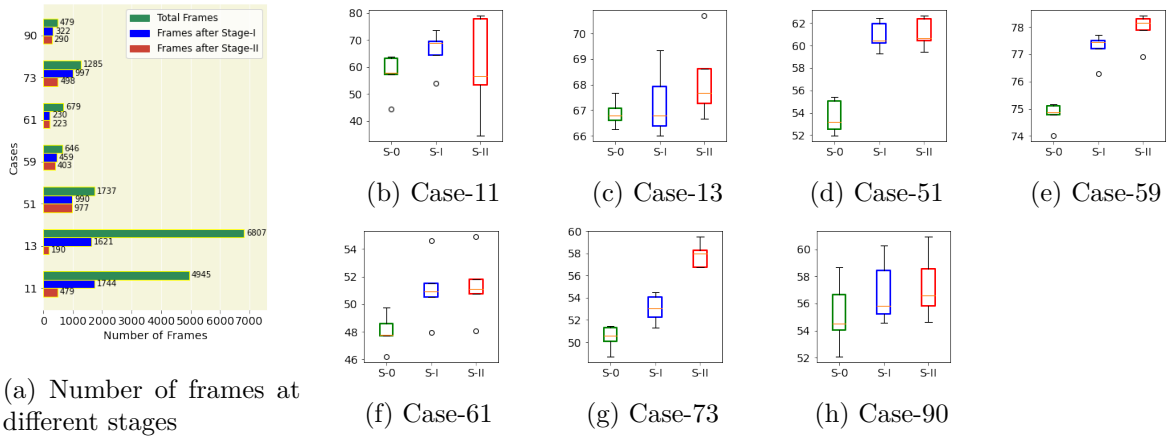
(m) Sequence-13 (n) Sequence-14 (o) Sequence-15 (p) Sequence-16 (q) Sequence-17 (r) Sequence-18



(s) Number of frames at different stages

Figure 3.9: The box plots (a)-(r) are associated with CVC-VideoClinicDB outcomes. The frames retained after different stages are represented by S-0: raw frames, S-I: frames obtained after Stage-I, S-II: frames obtained after Stage-II. The vertical axis represents the mAP@0.5:0.95. (s) shows the change in the number of frames at different stages. The dashed bars represent the updated frame count after CS tweaking.

trained our model and the baseline model using the SUN database and presented the testing results on the SUN test set and CVC-VideoClinicDB. The associated results are reported in AUPRC (Fig. 3.5 (i)-3.5 (p)) and Table 3.4. The table includes two types of scenarios, localization on test sets: a) before discarding non-polyp frames and b) after discarding non-polyp frames. In both scenarios, the proposed model achieved 2% and 5% increase in mAP@0.5:0.95 and an improvement of about 1% and 4% in F1-score when evaluated using the SUN test set and CVC-VideoClinicDB, respectively. Some sample images demonstrating the qualitative comparison of our model with the YOLOv5 are shown in Fig. 3.8. For a fair



(a) Number of frames at different stages

(b) Case-11 (c) Case-13 (d) Case-51 (e) Case-59

(f) Case-61 (g) Case-73 (h) Case-90

Figure 3.10: (a) shows the change in the number of frames at different stages. The box plots (b)-(h) demonstrate the performance of our proposed polyp localization model using frames of the SUN database obtained at different stages. The frames retained after different stages are represented by S-0: raw frames, S-I: frames obtained after Stage-I, S-II: frames obtained after Stage-II. The vertical axis represents the mAP@0.5:0.95.

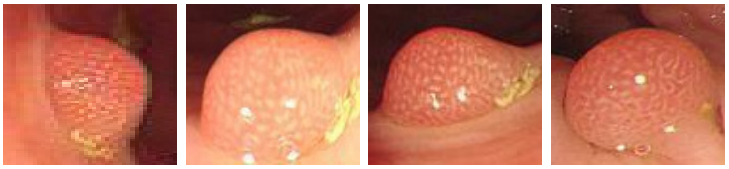
comparison, we have set the same confidence threshold of 0.5 while generating these sample images. The superior performance of our model is evident from the undetected polyp frame by YOLOv5. Moreover, our model predicts more accurate bounding boxes as compared to that of the YOLOv5.

Validating the significance of Stage-II: The results in Table 4.1 not only present the superior localization performance of our model but also indicate the effectiveness of Stage-II in the proposed pipeline. After discarding non-polyp frames, our model and the YOLOv5 achieved enhanced detection and localization in terms of precision, recall, F1-score, and mAP. A significant increase of about 2% and a slight improvement in mAP@0.5:0.95 can be observed in the case of the SUN database and the CVC-VideoClinicDB, respectively.

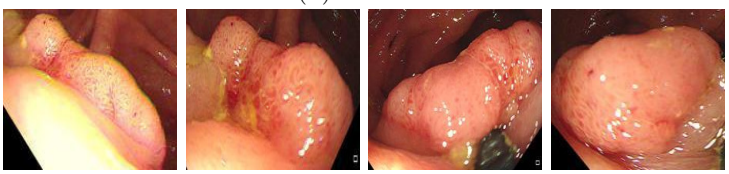
Case-wise Statistical Analysis

In real-world settings, the keyframe selection process is majorly patient-specific. Therefore, we analyzed our approach statistically using each case in the test sets independently. We executed our proposed localization model five times and assessed each of them using individual cases. The related box plots are shown in Fig. 3.10 and Fig. 3.9. The box plots present a comparative analysis of the localization performance on a) raw frames (S-0), b) frames obtained after Stage-I (high-quality frames; S-I), and c) frames obtained from Stage-II (only polyp frames; S-II) in terms of mAP@0.5:0.95. It is noteworthy that the results are not cherry-picked, and box plots of every case from the test sets (SUN database and

(a) Case-59



(b) Case-13



(c) Case-51

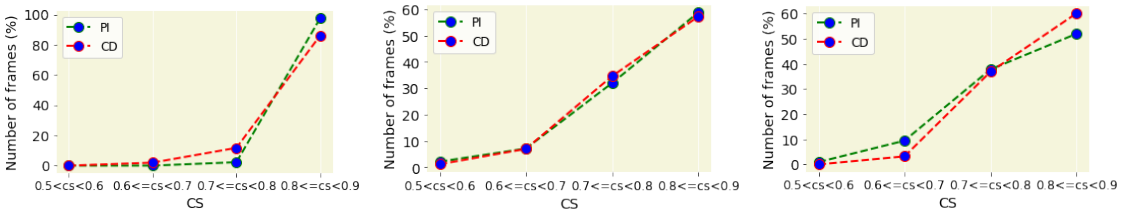
Figure 3.11: The figure shows the diverse view of polyps (ROIs) obtained from the final keyframes. Each row displays ROIs of a specific case.

CVC-VideoClinicDB) are included.

The results show that in 60% of the box plots, either the interquartile range or the median of S-I and S-II is present above that of the S-0. This signifies that in all such cases, S-I and S-II perform significantly better than S-0. In 20% of the cases, the boxes of S-0, S-I, and S-II overlap, representing insignificant differences, i.e., all are comparable. It is observed that S-I and S-II performed significantly superior or comparable to S-0 in all the cases of the SUN test set. The remaining 20% plots (5 cases from the CVC-VideoClinicDB) present better results with S-0. One of the reasons for the slightly low performance in the case of CVC-VideoClinicDB is the cross-dataset evaluation. Another important noticeable point is that out of these 5 cases, 3 are such cases (cases 5, 13, and 15) that required tweaking of the CS in Stage-I. This signifies that the video quality of these patients is relatively low. Still, our proposed attention-based localization model was able to identify polyps in low-quality frames when more such frames were presented to it in S-0. Nevertheless, the majority of the cases support the importance of Stage-I and Stage-II for keyframe selection.

Final Keyframes

Before this step, we discarded objects with CS lower than 0.5 to retain relatively more crucial frames that carry relevant polyp-characterizing features. This is followed by a score generation process based on the three criteria mentioned in Section 3.2.5. An example is



(a) Case-13 (b) Case-51 (c) Case-90

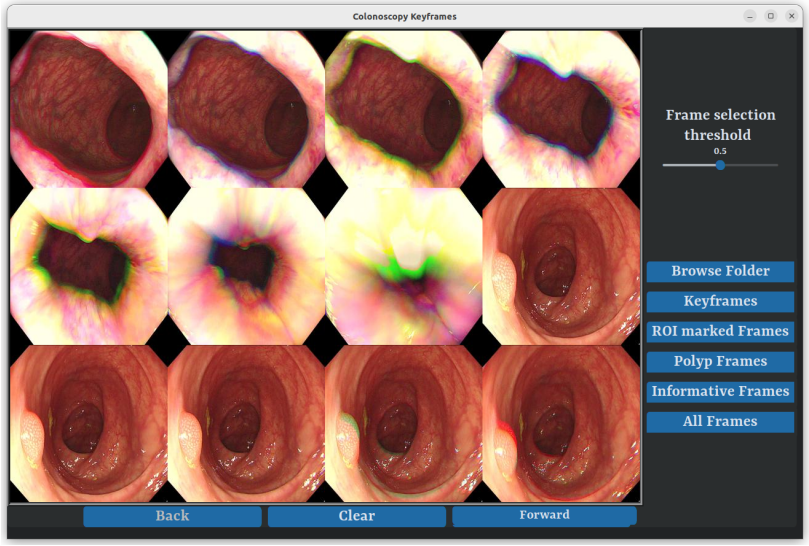
Figure 3.12: The graphs depict the relation between the CS, PI, and CD. Polyp frames with PI> PI_{avg} and $CD < CD_{avg}$ are likely to be detected with higher CSs.

shown in Fig. 3.7. The first sub-plot is obtained from Stage-III, and the next three are generated using Stage-IV outcomes. The objective of the dissimilarity scores is to form clusters, and the other three contribute to the selection of the most significant frame from each obtained cluster. To determine the keyframe from a group of redundant frames, we considered the highest CS followed by the highest PI and the lowest CD. The final scores thus generated are shown in the last sub-plot of Fig. 3.7. The frame with the highest final score within a cluster is chosen as the keyframe. As we get a collection of non-redundant representative frames at the end, diverse polyp views captured from different directions are obtained. Some samples are shown in Fig. 3.11.

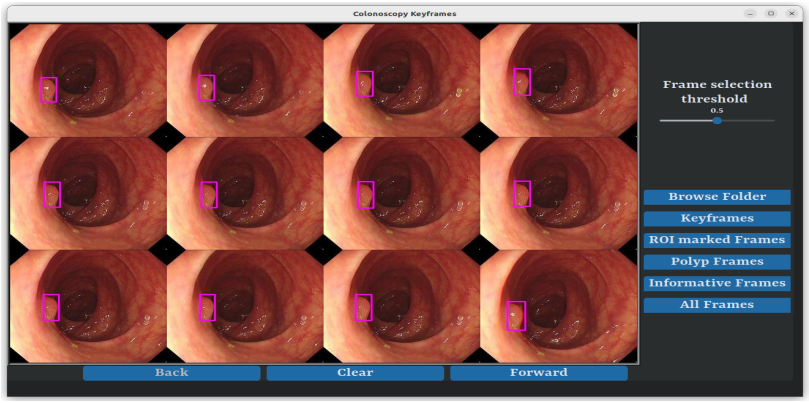
In this work, the results corresponding to each stage are validated by the subsequent stages, and in each case, we received positive outcomes. Similarly, to validate the effectiveness of the score-based approach used in the final stage, we investigated the general trend followed by the CSs, PIs, and CDs. We collected frames with CS> 0.5 and computed the average, PI_{avg} and CD_{avg}, over the PIs and CDs, respectively. We then counted the number of frames with $PI > PI_{avg}$ and $CD < CD_{avg}$. We observed that the number of such frames tends to increase with higher values of reported CSs. It implies that the automated localization model also detects polyps with high probability when they are centrally located and cover more image area. The related graphs of some cases are shown in Fig. 3.12.

Number of Keyframes and overall performance

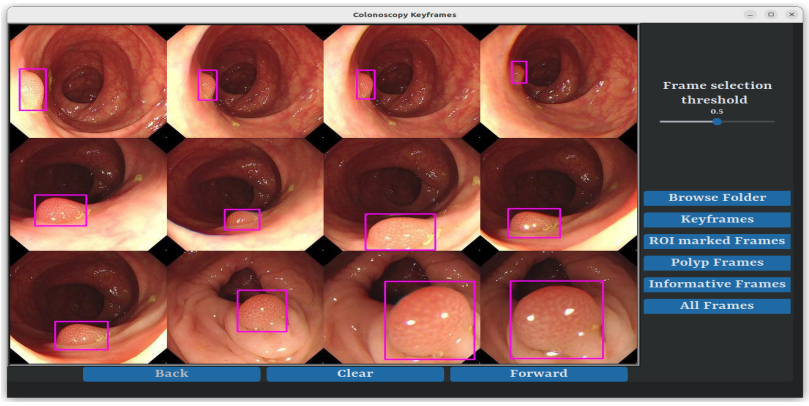
The final keyframes extracted by our complete pipeline present a reduction of 96.3% and 94.02% frames in the SUN database and the CVC-VideoClinicDB, respectively. These representative frames are capable of achieving enhanced or comparable diagnoses with low resource requirements. This could be validated from the intermediate stages' results in which we obtained improved localization with 81.54% and 40.89% decrease in the total number of frames of the SUN database and the CVC-VideoClinicDB, respectively. A similar



(a) All frames



(b) Informative frames with ROIs marked



(c) Keyframes with ROIs marked

Figure 3.13: Some screenshots of the proposed GUI. It consists of several options to easily navigate through different proposed stages.

trend is seen in the polyp detection outcomes. The change in the number of frames during the intermediate stages is shown in Fig. 3.10 (a) and Fig. 3.9 (s).

GUI for keyframes visualization and easy navigation

We designed a GUI to allow clinicians to navigate through the different stages of our proposed framework. It has three panes: the left pane displays frames, the right pane consists of different stage navigating options, and the bottom pane has buttons to scroll through the frames of a single video. The number of selected frames could be modified using a slider that changes the CS of the models. The bottom pane's *Back* and *Forward* buttons help switch to previously visited and yet-to-be-visited frames, respectively. The *Clear* button unloads the left pane. Screenshots of the GUI presenting visual results of some options are shown in Fig. 3.13. To the best of our knowledge, no GUI is designed in our research community that facilitates keyframe selection from colonoscopy videos with such a wide range of navigating options.

3.4 Limitations and Discussion

Although our proposed framework achieved promising outcomes with multiple validation checks, the approach requires human intervention for customized settings. Control of clinicians over some settings could be useful many times, but there is a trade-off between the benefits of clinician control and full automation. For example, in Stage-I, the proportion of frames to be preserved is handled manually and is decided on a predefined criterion. These settings are required for some patients' data in which not enough frames are retained. Such situations could arise during testing in real-world scenarios and are attributed to various factors. These factors include inadequate patient preparation, inappropriate recordings due to unexpected events, and limitations of the automated model in identifying informative frames from videos acquired using different imaging modalities. The last factor could be verified from the case-wise analysis in Stage-IV, which shows impressive results when training and testing are conducted on the same source domain using the SUN database. On the contrary, few failure cases are reported on cross-dataset evaluation.

The lingering gap due to the above limitations creates opportunities for a system with more automated settings and, more specifically, for a single-stage model that can integrate all capabilities present in our proposed multi-stage framework. Furthermore, domain adaptation could also be explored for better performance, thus reducing the need for frequent

tuning of settings. In the future, we will also conduct a small study on the users' experience with GUI and will incorporate the obtained feedback for improvement. Despite the few limitations, our work attains compelling results and could save clinicians from the burdensome task of manual frame selection.

3.5 Chapter Summary

In this chapter, we addressed the issue of the massive archival of colonoscopy videos and the huge amount of manual effort involved in reviewing these video streams. In this context, we proposed an automated system with a four-stage keyframe extraction pipeline. Our work adopted deep learning models that consider the image quality and select clinically significant data. We also reduced the correlated frames, providing diversity among keyframes along with a multi-view of polyps. In addition, our proposed multi-scale attention-based model has shown improved polyp localization outcomes with a 2% and 5% increase in mAP results using the SUN database and the CVC-VideoClinicDB, respectively. To validate the effectiveness of the proposed pipeline, we conducted extensive experiments with case-wise and cross-dataset evaluations. The results present the superiority of the proposed approach with enhanced polyp detection and localization. Furthermore, the detection processing time is lowered by 33.28% and 59.99%, and a substantial reduction of 96.3% and 94.02% in the total number of frames is achieved using the SUN database and the CVC-VideoClinicDB, respectively.

4

Artifacts Removal

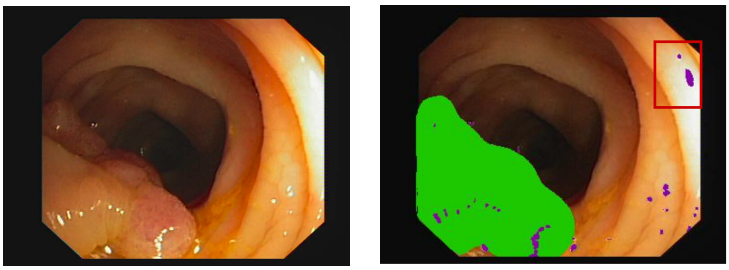
Chapter Highlights

- We observe that some patients' data required confidence score tuning and human interventions to obtain a sufficient number of keyframes for analysis.
- The primary reason behind the need for confidence score tuning is the inappropriate video recording with an enormous amount of uninformative frames.
- Such cases need to be handled appropriately, as keyframe selection alone may not provide sufficient information for a comprehensive diagnosis.
- We propose an adversarial network based solution to translate uninformative frames to clinically significant frames. Such translation helps uncover obscured clinical details.
- We further propose a specularity segmentation technique involving a DWT based encoder-decoder network.
- This chapter is based on the publications "*Can Adversarial Networks Make Uninformative Colonoscopy Video Frames Clinically Informative? (Student Abstract)*" presented at AAAI 2023 and "*A DWT-based encoder-decoder network for Specularity segmentation in colonoscopy images*" published in Multimedia Tools and Applications.

This chapter addresses the problem of artifacts that cover a large proportion of colonoscopy video data. As inappropriate video recordings could lead to an insufficient number of keyframes and need manual interventions, we propose two automated solutions to deal with such cases. In the first solution, we propose an adversarial network based framework to convert uninformative frames to clinically relevant frames. We examine the effectiveness of the proposed approach by evaluating the translated frames for polyp detection using YOLOv5. Preliminary results present improved detection performance along with elegant qualitative outcomes. We also examine the failure cases to determine the directions for the research community in the related domain. In the second solution, we perform specularity segmentation, which is a crucial pre-processing step for efficient computational diagnosis. The presence of these specular highlights could mislead the detectors that are intended to identify biomarkers precisely. Conventional methods adopted so far do not provide satisfactory results, especially in the overexposed regions. In this chapter, we propose a novel deep learning based approach that performs segmentation following a multi-resolution analysis. This is achieved by introducing DWT into the proposed model. We replace the standard pooling layers with DWTs, which helps preserve information and circumvent the effect of overexposed regions. All analytical experiments are performed using a publicly available benchmark dataset, and an F1-score (%) of 83.10 ± 0.14 *is obtained on the test set. The experimental results show that this technique outperforms state-of-the-art methods and performs significantly better in overexposed regions.*

4.1 Introduction

In the existing works, the well-trained models intended for polyp detection, segmentation, and classification still report limited diagnostic success. This limited success of automated methods is attributed to low-quality frames in the video samples, which contain various artifacts, namely, ghost colors, low-illumination, motion blur, interlacing, and fecal depositions (discussed in Chapter 1). Our keyframe selection approach deals with most such scenarios; however, cases involving inadequate patient preparation and images with specular highlights need specific handling. As stated in the previous chapter (Section 3.4, Chapter 1),

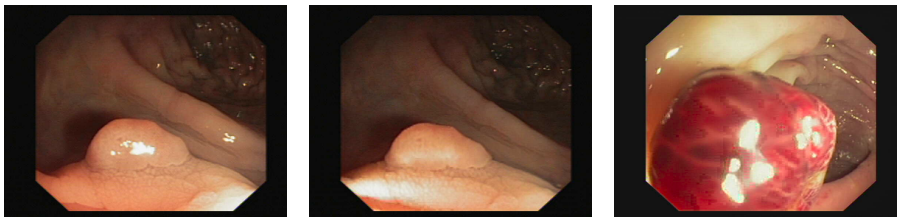


(a) (b)

Figure 4.1: Specular highlights in a colonoscopy image. (a) Original image, (b) Image representing specular highlights in purple. Polyp is shown in green with some highlights on its surface. Red frame shows specular reflections in overexposed regions.

inadequate patient preparation and low-quality video recordings due to unexpected events might need manual intervention to perform keyframe selection. Such scenarios signify that obtaining informative frames for appropriate diagnosis could be difficult at times. Like our approach to keyframe selection, some other related fields of laparoscopy and endoscopy followed keyframe selection [43] or performed super-resolution [124] to overcome similar issues, but no work in the colonoscopy domain explored the idea of extracting obscured clinical details from such low-quality uninformative video frames. Apart from the artifacts like ghost colors, low-illumination, motion blur, interlacing, and fecal depositions, the problem of specular highlights also needs to be addressed. Unlike other artifacts, specular highlights could be present in small regions even in visually clear and non-blurred images. Therefore, they are dealt with separately in this chapter. The basic details of other artifacts are already discussed in the previous chapter, while a concise overview of specular highlights is presented in this section.

Specular reflections are the bright spots that give a highly contrasted appearance in colonoscopy images. The watery shiny surface, also called mucosa, generates specular highlights when the colonoscope illuminates its surface in a perpendicular direction. These highlights distort the pixel values and yield erroneous outcomes. An instance with highlights on polyp's surface is shown in Fig. 4.1. Bernal *et al.* [10] illustrated the supporting role of specularity detection while executing the polyp detection algorithms. Such highlights mislead detectors, thus becoming the source of errors. According to Sanchez *et al.* [55], the presence of specular highlights could even lead to inaccurate polyp histology analysis. This could sometimes classify a non-adenomatous polyp as an adenomatous polyp, leading to incorrect histological results. Moreover, the performance of deep learning based cancer detectors majorly depends on the quality of data. That's why the pre-processing of



(a) (b) (c)

Figure 4.2: (a) Polyp with specular highlights, (b) Polyp (same as in (a)) without specular highlights, (c) Polyp's textural patterns partially corrupted by specular reflections

colonoscopy images is the first and most important step before feeding them to automated cancer detectors. Figure 4.1 presents an example with highlights in overexposed areas. In Fig. 4.2 (a) and Fig. 4.2(b), difference can be noticed between a polyp with specular highlights and the one without them. Fig. 4.2 (c) shows the obstruction caused by specular highlights in clearly visualizing the polyp's textural patterns. Considering the scenarios above, it becomes imperative to detect specular highlights in order to aid computational diagnostic methods. The related methods in the literature are dependent on thresholds and handcrafted features. Such approaches are liable to failures, and the results obtained so far are not satisfactory.

In this chapter, two solutions are presented:

- We investigate whether GANs can convert uninformative frames to informative frames. In this direction, we propose a GAN-based image-to-image translation approach to generate informative frames from the degraded frames of the colonoscopy videos.

- We propose a novel automated approach that aims at performing specularity segmentation in colonoscopy images. This method could assist clinicians to get better-preprocessed images and could lead to better performance of cancer detectors.

4.1.1 Contributions

The main contributions of this chapter are summarised below:

- **Solution-1: Adversarial Network to translate uninformative colonoscopy frames to clinically significant frames (Translation approach)**

  - To the best of our knowledge, this is the first framework to address the issue of uninformative colonoscopy frames using adversarial networks.

  - We investigate the impact of translating uninformative frames on polyp detection performance and discuss future directions in this context.

- **Solution-2: DWT-based encoder-decoder network for segmentation of specular highlights (Specularity segmentation approach)**

  - We propose a fully automated and efficient method to perform specularity segmentation in colonoscopy images. To the best of our knowledge, this is the first work that has explored deep learning architectures in this domain.

  - Our method outperforms the state-of-the-art methods and has shown satisfactory improvements over the limitations encountered in earlier works. These limitations include the failure to perform precise specular reflection segmentation in overexposed regions and inaccurate relative highlight pixel detection.

  - The proposed approach combines hierarchical scale-based Wavelet Packet Transform (WPT) with encoder-decoder architecture. This assists in extracting multi-resolution-based features. Utilizing the concept of DWT in place of pooling layers helps preserve image contextual and spatial information.

  - Our work proposes a change in the standard low-frequency sub-band of DWT, and this has shown significant improvement in segmentation results in overexposed areas. This improvement is supported by an ablation study for thorough analysis.

4.2 Solution-1: Translation Approach

The overview of the proposed framework is shown in Fig. 4.3. Given the uninformative colonoscopy frames $\{\mathfrak{a}_i\}_{i=1}^{M}$ from domain $\mathcal{A}$, the aim is to learn a mapping function $G_{AB} : \mathcal{A} \rightarrow \mathcal{B}$ to generate frames such that the data distribution of obtained frames is indistinguishable from that of informative colonoscopy frames $\{\mathfrak{b}_j\}_{j=1}^{M'}$ of domain $\mathcal{B}$. Due to the unavailability of paired data, our work is inspired by the unpaired translation approach of CycleGAN [125]. Hence, another mapping function $G_{BA} : \mathcal{B} \rightarrow \mathcal{A}$ is also introduced. Our implementation involves ResNet-based generators and PatchGAN discriminators D_A and D_B. The CycleGAN objective integrates adversarial loss and cycle-consistency loss.

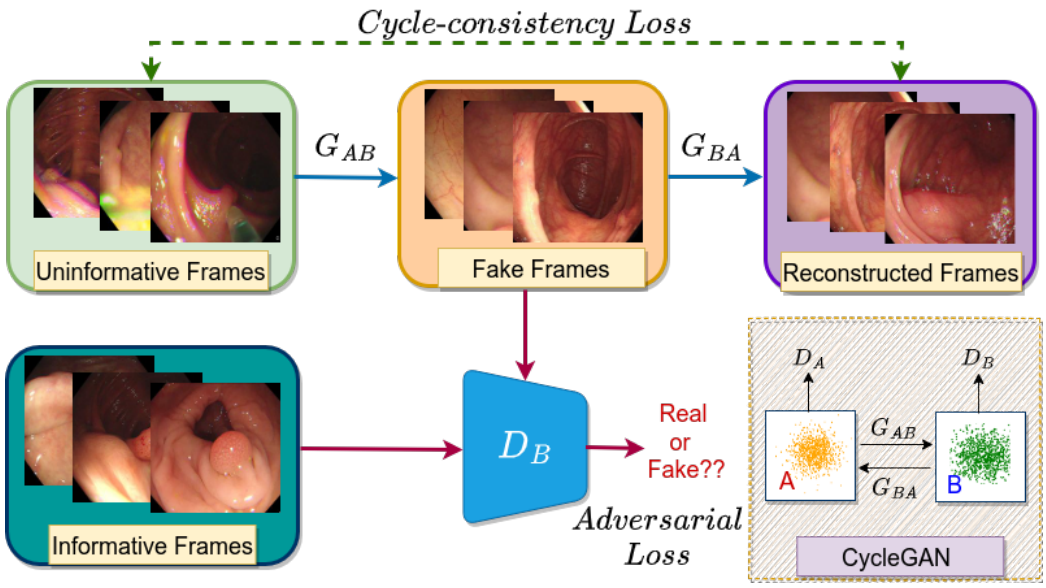


Figure 4.3: The proposed framework contains two generators G_{AB} and G_{BA} and two discriminators D_A and D_B.

The adversarial loss can be expressed as:

$$\begin{aligned} L_{adv}(G_{AB}, D_B) = \mathbb{E}_{\mathfrak{b}\sim p_{data}(\mathfrak{b})}[(D_B(\mathfrak{b}) - 1)^2] \\ + \mathbb{E}_{\mathfrak{a}\sim p_{data}(\mathfrak{a})}[(D_B(G_{AB}(\mathfrak{a})))^2] \end{aligned} \tag{4.1}$$

G_{AB} aims to translate uninformative frames such that they appear similar to the informative frames, while D_B tries to distinguish the translated frames from the high-quality, informative frames of domain B. In other words, D_B is trained to minimize $L_{adv}(G_{AB}, D_B)$ and G_{AB} is trained to minimize $\mathbb{E}_{\mathfrak{a}\sim p_{data}(\mathfrak{a})}[(D_B(G_{AB}(\mathfrak{a})) - 1)^2]$.

To ensure cycle-consistency and to reduce randomness in mapping, a cycle-consistency loss is used, which is given by:

$$\begin{aligned} L_{cyc}(G_{AB}, G_{BA}) = \mathbb{E}_{\mathfrak{a}\sim p_{data}(\mathfrak{a})}[\|G_{BA}(G_{AB}(\mathfrak{a})) - \mathfrak{a}\|_1] \\ + \mathbb{E}_{\mathfrak{b}\sim p_{data}(\mathfrak{b})}[\|G_{AB}(G_{BA}(\mathfrak{b})) - \mathfrak{b}\|_1] \end{aligned} \tag{4.2}$$

An identity mapping loss is also added to help preserve color in translated images. With this model, we intend to determine the clinically relevant details obscured by the artifacts. Furthermore, we carried out the following investigations:

1. Polyp detection is performed using YOLOv5 [126] to determine the impact of GAN-translated frames.

Metrics	SUN Database	
	Raw Frames	**Translated Frames**
Precision (%)	92.03±0.60	**93±0.87**
Recall (%)	88.9±3.12	**90.2±1.3**
F1-score (%)	90.4±1.51	**91.57±0.38**
mAP@0.5 (%)	95.37±0.95	**95.6±0.21**
mAP@0.5:0.95 (%)	**57.53±0.32**	57.07±0.31

Table 4.1: Comparative analysis of polyp detection results

2. Qualitative analysis is done to identify the artifacts successfully handled by the CycleGAN and analyze the ones that still persist in the translated frames.

4.2.1 Experiments

Dataset and Training Details

To assess the effectiveness of the adversarial approach in mitigating the impact of artifacts, we conducted experiments using a publicly available SUN database [95] consisting of $1,09,554$ non-polyp and $49,136$ polyp frames. In addition to the localization information, the polyp frames are manually annotated by experts as informative or uninformative. We used only the polyp frames with a patient-wise split. The translation is done on a Titan Xp GPU at 14 frames per second.

Performance Evaluation

We report the results based on visual perception and consider feature space representation by evaluating the polyp detection outcomes using YOLOv5. We conducted training and testing in two scenarios using: a) Raw frames comprising both high and low-quality frames and b) Translated frames along with high-quality frames. The results in Table 4.1 show that the translated frames complement the detection ability of YOLOv5 in terms of precision, recall, F1-score, and mAP@0.5. The detector correctly identified more polyps with lower deviations, presenting a more robust model. However, this is achieved with slightly less precise bounding boxes, as indicated by a minor decrease in mAP@0.5:0.95. Fecal depositions, ghost colors, and low-illumination are significantly reduced using CycleGAN, as shown in Fig. 4.4. However, motion blur and interlacing are not handled adequately in the process. This could be overcome by adopting blur removal approaches.

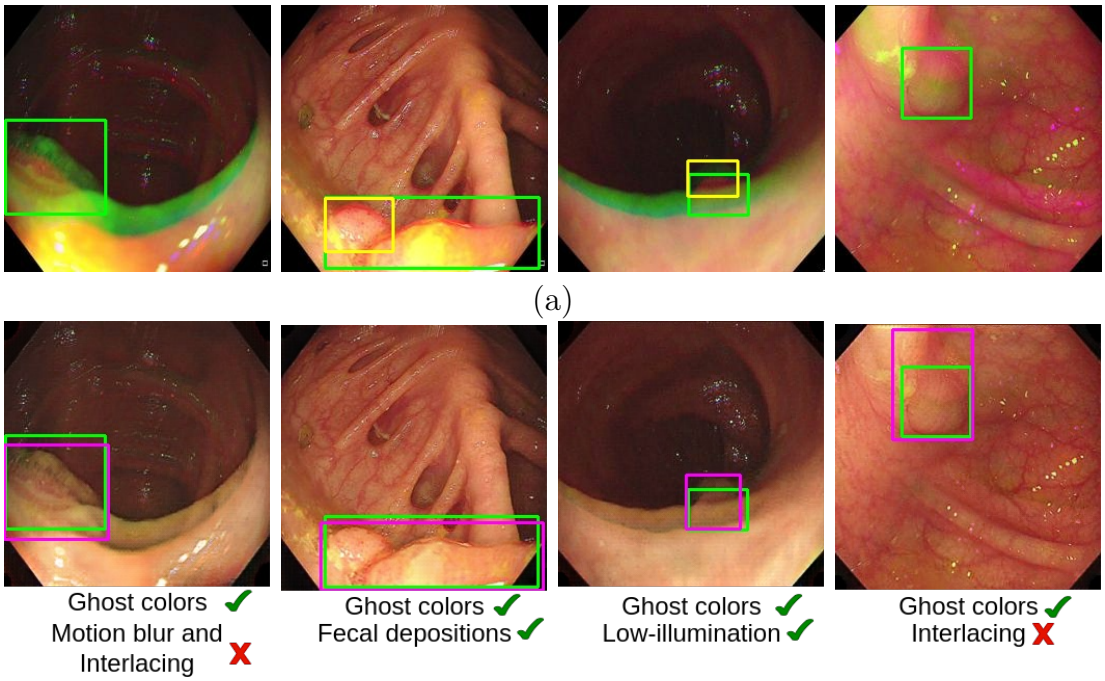


Figure 4.4: Detection performance using: (a) Raw frames and (b) Translated frames. Green bounding boxes denote the ground truth. Ticks and cross marks represent the successful and unsuccessful artifacts translations, respectively.

4.3 Solution-2: Specularity Segmentation Approach

Fig. 4.5 provides an overall idea of the steps involved. Inspired by the competent performance of U-Net[66] and other diverse sets of deep learning architectures based on it [127–130], we adopted the encoder-decoder architecture of U-Net as the basic structure of our proposed model. Further improvements have been incorporated to make it more suitable for the experimental requirements related to our problem definition. The encoder takes the blue channel of the given image as input and extracts the corresponding features. These features are passed to the decoder that performs reconstruction and provides a binary image with segmented specular highlights. A detailed explanation of the complete architecture, with reasoning, is provided in the subsequent sections. A summary of the methodology followed in our work has been depicted in Fig. 4.6.

While using deep learning models, the feature maps are generally progressively downsampled via max-pooling to attain a large valid receptive field. This is crucial to capture enough information to make correct decisions, but this compromises image resolution. As the resolution decreases with downsampling, information pertaining to small objects gets

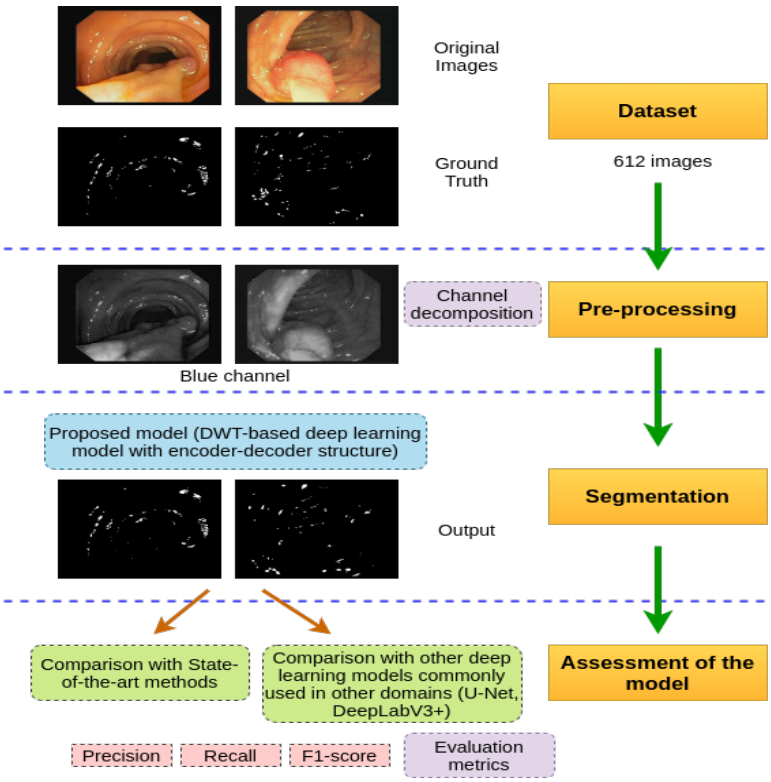


Figure 4.5: Flowchart depicting the steps followed in the proposed work

exhausted. Thus, the coarse features obtained from deep layers could miss small objects' cues, which could be difficult to recover even by using skip connections [131]. As stated in [55], most of the specular regions in colonoscopy images are confined to only a few pixels, making these regions very small in size. To make such regions discernible and prevent distortion of spatial acuity, a specific method is required to increase the receptive field without losing the resolution. Another problem to be considered is the presence of overexposed regions in colonoscopy images. Overexposed regions have not been categorized as specular highlights by experts. They are not eligible as per the conditions [55] specified for a region to be considered specular. However, these regions impose a negative impact on the performance of specularity detection algorithms.

Considering the limitations of the conventional architecture and the requirements of the specified problem, the proposed model is intended to preserve maximal information during feature extraction. At the same time, it adopts a method to circumvent the effect of overexposed regions, up-to some extent. Our model replaces pooling layers with multi-level Haar wavelet and hence, combines spectral analysis with CNNs. Moreover, the frequency and localization properties of DWT help in extracting discriminative features in

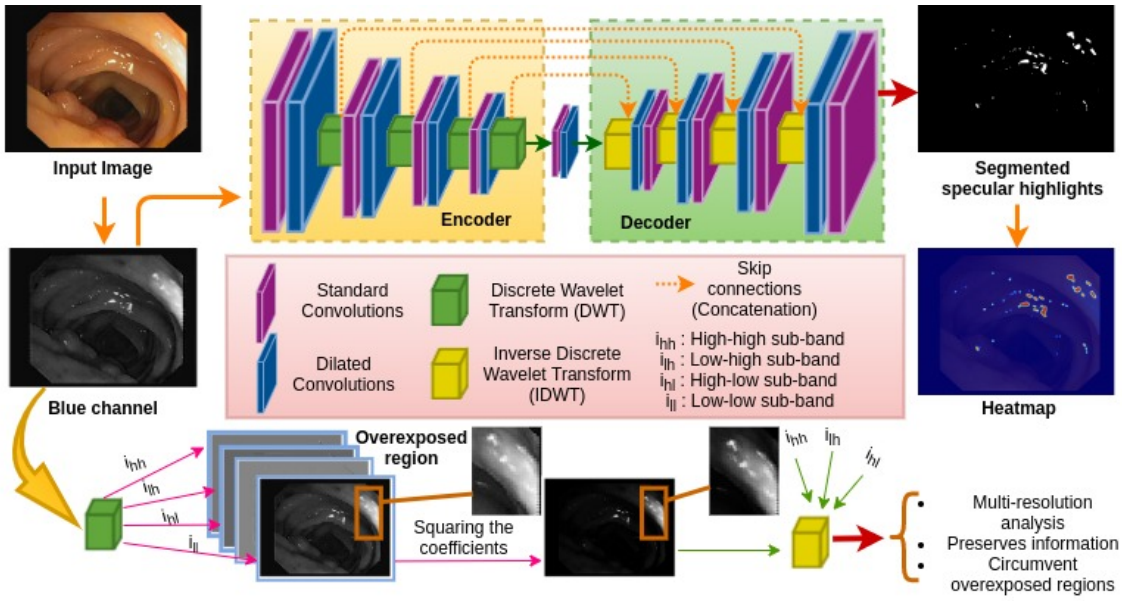


Figure 4.6: Summary of the methodology followed for the segmentation of specular highlights

the frequency domain, at hierarchical scale levels, while preserving image details.

4.3.1 Problem Definition

Given a colonoscopy image set Q, our motive is to assign each pixel of an image to a class belonging to either specular regions or non-specular regions. We have $Q = \{I_i, O_i\}_{i=1}^{m}$, where I_i represents original RGB images, and $O_i = \{o_j^{(i)}, j = 1, 2, ..., |I_i|, o_j^{(i)} \in \{0, 1\}\}$ denotes their corresponding ground truth results/masks. Every mask has each pixel labelled as $o_j^{(i)} = 1$ for representing specular regions and $o_j^{(i)} = 0$ for representing non-specular regions.

In [55], it is assumed that the diffuse component is likely to appear reddish in colonoscopy images; thus, the specular component could be better observed in the blue channel. Considering this observation, we have used blue channel I_i^{blue} of the corresponding original image I_i as our model's input. The aim is to train a pixel-wise classifier to learn the following mapping function:

$$\hat{O}_i = M_{wspec}(I_i^{blue}) \tag{4.3}$$

where M_{wspec} represents our proposed model and $\hat{O}_i$ is the segmented output.

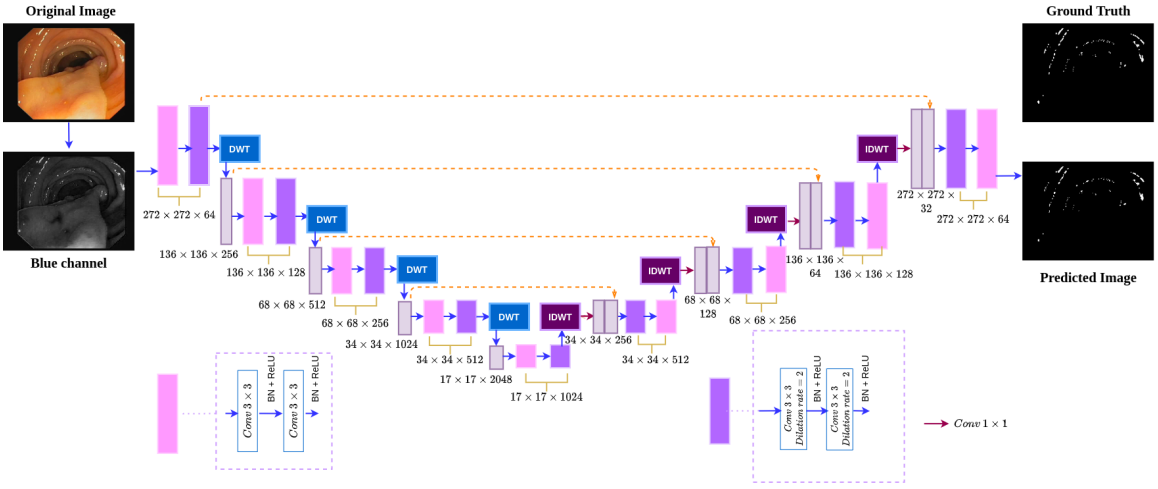


Figure 4.7: Detailed architecture of the proposed model. The model consists of DWTs and IDWTs at each level of encoder and decoder, respectively.

4.3.2 Proposed Architecture

As shown in Fig. 4.7, the backbone architecture used is based on encoder-decoder structure, and the inclusion of DWT is inspired by the model MWCNN [132]. As per the reason mentioned in Section 4.3.1, the original images' blue channel is provided as input to the model. Our model has a total of 36 layers. Every individual CNN block comprises 4 convolutional layers, each with 3×3 filter size, followed by BN and ReLU. Among the 4 convolutional layers, the two perform the dilated convolutions [133] with a rate of 2. Though DWT is expected to be effective in increasing the receptive field, introducing dilation in a few convolutional layers in our model prevents the missing of small crowded specular regions. Small objects in an image carry little signals, which could be lost in the later layers. For accurate detection of such small regions, especially those present in the close grouping, the most crucial factor to be considered is the context [134]. As most of the specular regions in colonoscopy images are very small in size and are crowded as well, adopting dilation along with DWT has shown better results.

On the encoder side, a DWT is placed as a link between two CNN blocks. The low-frequency sub-band is first processed to subdue the effect of overexposed regions. The related details are mentioned further in this section. After this processing, all the sub-bands are taken as input to the next CNN block. According to Li *et al.* [127], the sub-bands obtained after DWT are dependent, and this dependency should not be ignored for the purpose of attaining adequate restoration results. After applying DWT, the number of

feature maps increases, and the immediate next convolutional layer reduces these feature maps to provide a compact representation of features to the subsequent layers. On the decoder side, IDWT is deployed between two consecutive CNN blocks. The feature maps obtained after applying DWT are directly concatenated with the IDWT output presented one level below in the expanding sub-network, using skip connections. The last convolution performed is not followed by BN or ReLU.

In 2D DWT, a given image I_i^{blue} is decomposed into four sub-band images, namely, low-low ($\mathfrak{i}_{ll}$), high-low ($\mathfrak{i}_{hl}$), low-high ($\mathfrak{i}_{lh}$) and high-high ($\mathfrak{i}_{hh}$). For brevity and clarity, I_i^{blue} is replaced with a Greek $\mathfrak{i}$ at some places. The decomposition process is illustrated in Fig. 4.6. The low-frequency component, $\mathfrak{i}_{ll}$, also known as approximation coefficients, contains the average information of I_i^{blue} and other high-frequency components or the detail coefficients, $\mathfrak{i}_{hl}$, $\mathfrak{i}_{lh}$ and $\mathfrak{i}_{hh}$ constitute the edge details, thus provide a refined image. This DWT procedure could be considered as a convolution operation between the image $\mathfrak{i}$ and four filters $\mathfrak{f}_{ll}$, $\mathfrak{f}_{hl}$, $\mathfrak{f}_{lh}$ and $\mathfrak{f}_{hh}$, which is followed by downsampling. This can be represented by the given equation:

$$\mathfrak{i}_s^{(k)} = (\mathfrak{i} \otimes \mathfrak{f}_s) \downarrow_2 \tag{4.4}$$

where $s \in \{ll, hl, lh, hh\}$. This process is similar to the pooling operation up to some extent. The downside of pooling lies in the information loss encountered during down-sampling, which negatively affects the reconstruction process. In the case of DWT, even though downsampling is performed, the biorthogonal property of DWT makes it possible to accurately reconstruct the image by using inverse discrete wavelet transform (IDWT) as:

$$\mathfrak{i} = IDWT(\mathfrak{i}_{ll}, \mathfrak{i}_{hl}, \mathfrak{i}_{lh}, \mathfrak{i}_{hh}) \tag{4.5}$$

During reconstruction, upsampling is performed over the four sub-bands, and then the corresponding filters are deployed for convolution on the upsampled output. This provides us with a segmented image of the same spatial size as that of the input image.

The above-mentioned procedure could be extended to multi-level WPT. For this, the previously obtained sub-bands are further decomposed using DWT. In case of two-level WPT, each of the four sub-bands obtained in the first-level is decomposed into another four sub-bands. This process could be carried out recursively in a similar manner to get higher levels of WPT.

The 2D DWT used in the proposed model is specifically focused on the Haar wavelet due to its simplicity and satisfactory speed performance. The four filters used in Haar

Wavelet can be defined as:

$$\begin{aligned} \mathfrak{f}_{ll} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \mathfrak{f}_{hl} = \begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix}, \\ \mathfrak{f}_{lh} = \begin{bmatrix} -1 & -1 \\ 1 & 1 \end{bmatrix}, \mathfrak{f}_{hh} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \end{aligned} \tag{4.6}$$

Low-frequency and high-frequency sub-bands of an image I_i^{blue} can be obtained by convolving the above filters with it. $(p', q')^{th}$ value of low-frequency component $\mathfrak{i}_{ll}$ can be computed as:

$$\begin{aligned} \mathfrak{i}_{ll}(p', q') = \mathfrak{i}(2p' - 1, 2q' - 1) + \mathfrak{i}(2p' - 1, 2q') + \\ \mathfrak{i}(2p', 2q' - 1) + \mathfrak{i}(2p', 2q') \end{aligned} \tag{4.7}$$

The $\mathfrak{i}_{ll}$ sub-band contains the main base information. When applied to a given image, the low-pass filter, $\mathfrak{f}_{ll}$, determines the averages of pairs of nearby values; hence, smooth variations in the image are portrayed by the approximation coefficients present in $\mathfrak{i}_{ll}$. Over-exposed regions are also visible in the same sub-band. To overcome these regions, each value in the component, $\mathfrak{i}_{ll}$, is squared, as shown in Eq. 4.8. Squaring each coefficient in this component almost maintains the relationship among pixels with similar values, but at the same time, it increases the difference among dissimilar nearby pixels. Due to this, the specular highlights show a greater difference relative to the overexposed regions. The result of this processing is shown in Fig. 4.8

$$\hat{\mathfrak{i}}_{ll} = \mathfrak{i}_{ll} .\times \mathfrak{i}_{ll} \tag{4.8}$$

where $.\times$ is the element-wise multiplication. Similar to Eq. 4.7, high-frequency components $\mathfrak{i}_{hl}$, $\mathfrak{i}_{lh}$ and $\mathfrak{i}_{hh}$, can also be computed. These components are comprised of sharp variations, and as a result, abrupt intensity changes on borders of specular highlights are easily captured by these sub-bands. For integrating all the sub-bands in the proposed model, they are concatenated as depicted by Eq. 4.9. This concatenated output is made to proceed to subsequent layers.

$$\mathfrak{i}_{sub\_cat} = \hat{\mathfrak{i}}_{ll} \oplus \mathfrak{i}_{hl} \oplus \mathfrak{i}_{lh} \oplus \mathfrak{i}_{hh} \tag{4.9}$$

Although DWT helps in the expansion of the receptive field, the small-sized, crowded specular regions are liable to vanish in the subsequent layers. According to Zhou *et al.*

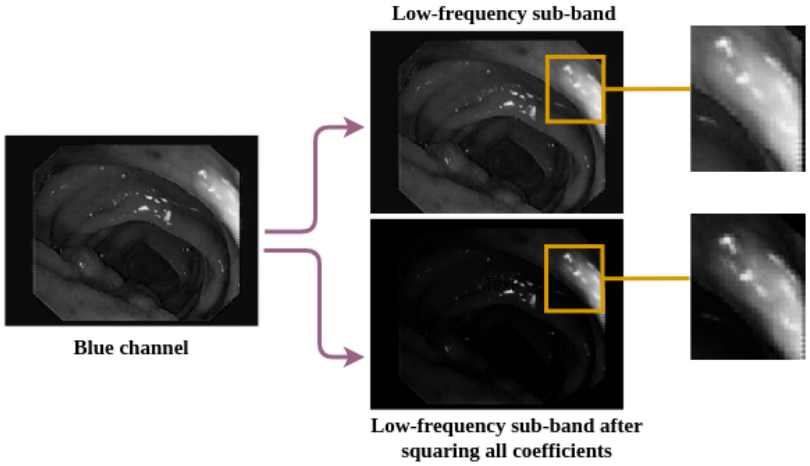


Figure 4.8: Suppressed overexposed regions obtained by squaring the Low-frequency sub-band coefficients

[135], the receptive field's true size is much smaller than the theoretical size, mainly in the last layers. To solve this problem, Yu and Koltun [133] proposed dilated convolutions using which the receptive field increases exponentially while the number of parameters follows a linear growth.

When compared to the conventional encoder-decoder structure, the proposed model replaces the pooling layers in the encoder part with DWTs. Accordingly, the transposed convolutions in the decoder part are replaced with IDWTs, ensuring accurate image reconstruction. Each level of transform is followed by CNN blocks. Every CNN block gets all the sub-bands together as input, which helps in maintaining inter-sub-band dependency. Moreover, DWT's frequency and localization properties are supposed to preserve textural and structural details, which is another advantage over the usual pooling layers. Another important benefit of using DWT in our model is its ability to handle overexposed regions in images. Many research studies have explored the role of DWT in image domain in the context of correcting overexposed areas as well as in other image enhancement-related tasks [136, 137]. In this work, we experimented with DWT in deep learning models to investigate the impact of Eq. 4.8 in overexposed regions. The obtained results provide improved specular detection performance in such areas. The evidence of the same is provided in Section 4.3.4 (Fig. 4.12, Table 4.4).

In our work, DWT has been embedded into the model to preserve specularity-relevant contextual and spatial information. The low and high-frequency sub-bands together form the same frequency content as that of the input colonoscopy images. In the decomposition

process, DWT also helps to increase the data, similar to data augmentation, which enhances the training process. Besides this, DWT can increase the receptive field of CNNs, hence making it easier to detect small specular highlights. The dilated filtering and pooling operations could be interpreted as closely related to DWT [132]. A general average pooling operation with a factor of 2 on an image I_i^{blue} is defined as:

$$\mathfrak{i}_{pooling}(p', q') = \frac{\mathfrak{i}(2p'-1, 2q-1) + \mathfrak{i}(2p'-1, 2q') + \mathfrak{i}(2p', 2q'-1) + \mathfrak{i}(2p', 2q')}{4} \tag{4.10}$$

which is similar to Eq. 4.4 except a fixed constant 1/4. This implies that decomposing images into sub-bands using DWT is related to dilated filtering.

4.3.3 Loss Function and Evaluation Metrics

Let $\{I_i, O_i\}_{i=1}^{TN} \subset Q$ be the training input-target pairs, where TN is the number of training samples used. $o_n \in O_i$ is associated with ground truth pixel labels representing the two classes and has value in {0,1}. The predicted probabilities of the two classes can be represented by $\mathfrak{p}_n \in [0, 1]$. The loss function used is a weighted sum of Dice Loss (DL) and BCE loss with logits. Reasons for using these loss functions include wide use of BCE in binary segmentation problems and superior performance shown by DL in case of imbalanced datasets, similar to what we are dealing with. In our dataset, the pixels of non-specular regions are much higher than those of specular regions. To tackle this issue, dice loss is used, as can be described by the given equation:

$$DL = 1 - \frac{2\sum_n \mathfrak{p}_n o_n + \gamma'}{\sum_n \mathfrak{p}_n + \sum_n o_n + \gamma'} \tag{4.11}$$

where γ' is a factor added for smoothing purposes and is initialized as $\gamma' = 1.0$

We have used three metrics for comparing our segmentation results with state-of-the-art methods. These metrics are precision, recall, and F1-score. As the negative samples in our case are much higher than the positive samples, using accuracy as the metric is inappropriate. Precision computes the number of true positives out of the samples predicted as positive; hence, a high number of negative samples does not affect the precision results. High precision leads to high performance towards accurate highlight detection. On the other hand, recall accounts for the number of positive samples predicted correctly out of the total positive samples presented by the ground truth. In this way, missed positive pixels are indicated. A high recall rate ensures the detection of more specular pixels. F1-score captures both precision and recall values and combines them into a single score.

4.3.4 Experiments and Results

Dataset and Training Details

The proposed model is evaluated on a publicly available dataset, CVC-ClinicSpec [55]. It contains 612 annotated specular highlight images related to colonoscopy. The ground truth of each image is labelled manually by experts. 57.43% of the specular regions are quite small, and their size ranges from 1 to 10 pixels. Only 3.70% of the specular regions are more than 100 pixels in size and can be considered large-sized specular highlights. To train our model, a training set of 367 images is constructed from the available 612 images of the CVC-ClinicSpec Dataset. From the remaining images, 123 are used as the Validation set, and 122 images are reserved for testing. The model is fed with resized images of size 272 $\times$ 272. During processing, the proposed model uses only the blue channel of the original images. Implementation and experiments are performed using Pytorch 1.4.0. Training is carried out on TITAN Xp GPU. We have used the SGD optimizer with momentum set to 0.9 in combination with the ReduceLROnPlateau scheduler. The batch size used is 10, and our model converges after 60 epochs. The model is initialized with a learning rate of 0.1 which is decayed by the scheduler in case the validation loss increases for 4 consecutive epochs.

Ablation Study

To study the role of different architecture segments, we have performed an ablation study by introducing them gradually with each new step. The results can be visualized in the graph shown in Fig. 4.9.

1) Encoder-Decoder + DWT in original form & IDWT: First, we considered the basic encoder-decoder architecture. It consists of DWT instead of pooling layers and IDWT instead of the usual upsampling or transposed convolutions. Initially, the low-frequency components are considered in their original form without squaring them. The F1-score obtained in this case is 82.17%.

2) Encoder-Decoder + DWT with $\hat{i}_{ll}$ & IDWT: To test the effect of the changes made in the low-frequency components of DWT, we further replaced the normal DWT functions with the updated DWT as defined in Section 4.3.2. This modification raised the F1-score by 0.59% and recall by 1.01%.

3) Encoder-Decoder + DWT with $\hat{i}_{ll}$ & IDWT + 2 dilated convolutions: To analyze the performance of dilation in our model, we changed two of the standard convolutions to

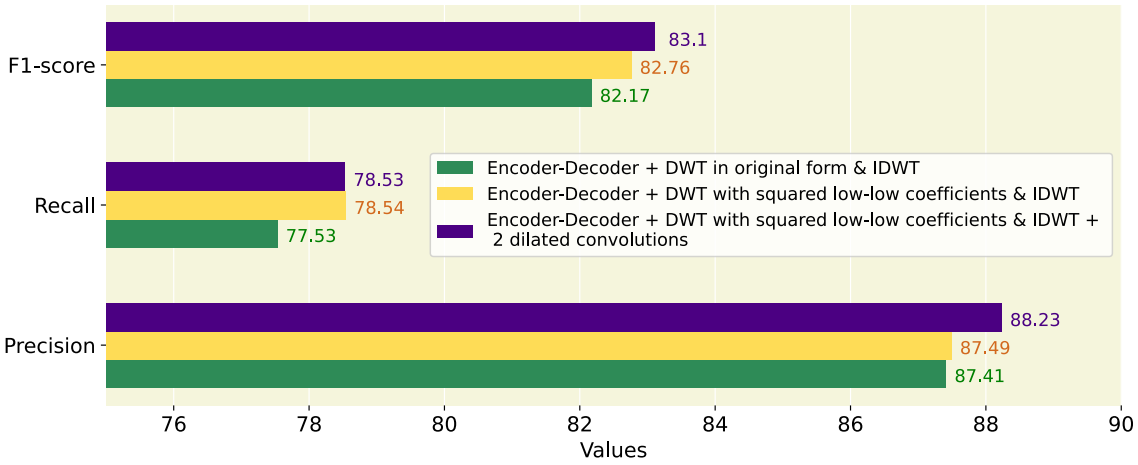


Figure 4.9: Ablation study related to the proposed model on the basis of precision, recall and F1-score

Table 4.2: Comparison with State-of-the-art methods

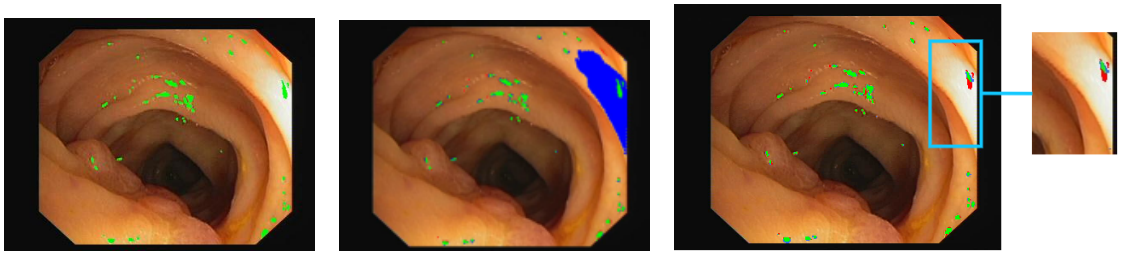
Method	**Precision (%)**	**Recall (%)**	**F1-score(%)**
BSSC [55]	72.33	78.89	75.47
RPCA [53]	34.38	-	-
Proposed	$\mathbf{88.23 \pm 0.58}$	78.53 ± 0.68	$\mathbf{83.10 \pm 0.14}$

dilated convolution with rate 2. This improved the F1-score further by 0.34% and precision by 0.74%.

Comparison with State-of-the-Art Methods

Comparative results of specularity segmentation are presented in Table 4.2. The results are based on the pixel-wise analysis, which takes into account the number of pixels correctly or wrongly detected by the algorithm. The segmentation results indicate that our proposed method has outperformed the state-of-the-art results in terms of F1-score and precision. The recall results obtained are also comparable. The methods considered for comparison are those which have particularly focused on colonoscopy images. This ensures that the analysis is done on the images acquired in similar environmental conditions and problems encountered by all the methods are also identical.

Sanchez *et al.* [55] discussed the impact of overexposed regions on the specularity segmentation algorithms. To analyze this impact, they excluded the images with overexposed regions and performed a validation experiment. Their results showed an improvement of



(a) Original (b) BSSC[55] (c) Proposed

Figure 4.10: Comparison of the proposed method with BSSC [55]: True positive pixels, false positive pixels and false negative pixels are represented in green, blue and red color, respectively

8.30% in precision, 2.14% in recall, and 5.36% in F1-score. This clearly indicates the negative impact that overexposed regions impose on the correctness of segmentation results. Considering such high increase in precision, it can be inferred that their method experienced difficulty in detecting specular highlights which are present within the overexposed regions. This makes the images more susceptible to false positives, as can be seen in Fig. 4.10 (b).

Li *et al.* [53] considered two divisions of highlight pixels: absolute and relative. Their work has mainly focused on detecting absolute pixels, while the latter, being more complicated to detect, are not appropriately handled. It has not strictly followed the ground truth of the CVC-ClinicSpec database, prepared by the experts. Contrary to the available ground truths, the overexposed areas are also included in the category of highlights and are segmented by their algorithm.

In our proposed method, we have considered the fact that computational diagnosis procedures in the colonoscopy domain rely on texture and color features, as stated in [138, 139]. More false positives due to overexposed areas can lead to loss of texture and color-related information. By dealing with such areas, false positives could be reduced during specularity detection. Later, these overexposed regions could be processed with some enhancement techniques to obtain informative features. The comparison shown in Fig. 4.10 indicates better performance of our model in the overexposed regions as it fetches less false positive pixels compared to BSSC.

Comparative Study with Some Deep Learning Models

The proposed work has not directly adopted the conventional encoder-decoder architecture but has made appropriate changes to the architecture according to the aforementioned objectives. To analyze our model's performance with respect to some other deep learning models which are commonly used in other domains related to semantic segmentation, we

Table 4.3: Comparative results with other deep learning models. The mean values of the experimental results along with the standard deviation are reported

Method	**Precision (%)**	**Recall (%)**	**F1-score (%)**
DeeplabV3+ [140]	73.48 ± 4.7	69.24 ± 3.13	71.28 ± 3.77
U-Net [66]	88.25 ± 1.32	76.58 ± 1.5	81.99 ± 0.92
Proposed	$\mathbf{88.23 \pm 0.58}$	$\mathbf{78.53 \pm 0.68}$	$\mathbf{83.10 \pm 0.14}$

have done some comparative study, which is presented below.

Models' Details and Training Settings

To compare the performance of our model with other models, we have considered two different architectures. The first one is U-Net, and the other one is DeepLabV3+ [140]. The motive behind choosing U-Net is mentioned in Section 4.3. The same can be justified for DeepLabV3+ by considering its adequate performance in semantic segmentation [141, 142] and its ability to extract features at arbitrary resolutions using atrous separable convolutions. In DeepLabV3+, Modified Aligned Xception is used as the backbone. To maintain uniformity and based on some prior experiments with the dataset, we have used the same hyper-parameters values as mentioned in Section 4.3.4. Both models are provided directly with the original images. As per the model's architecture and the paddings used, the input images are resized to 280×280 for both U-Net and DeepLabV3+.

Experimental Results

As shown in Table 4.3, our model has outperformed other chosen deep learning models in terms of precision, recall, and F1-score. The heatmaps shown in Fig. 4.11 provide an idea about the segmentation performance of different models. As observed from the respective images, DeepLabV3+ has performed better in some overexposed areas as compared to U-Net, but at the same time, it has missed some small specular regions. Few closely grouped areas are over-segmented and are merged into one region, ignoring their distinct boundaries. Considering the heatmaps associated with U-Net, it can be noticed that though more small highlights are detected, there are many false positives encountered in overexposed regions. The proposed model is observed to perform better than both the models.

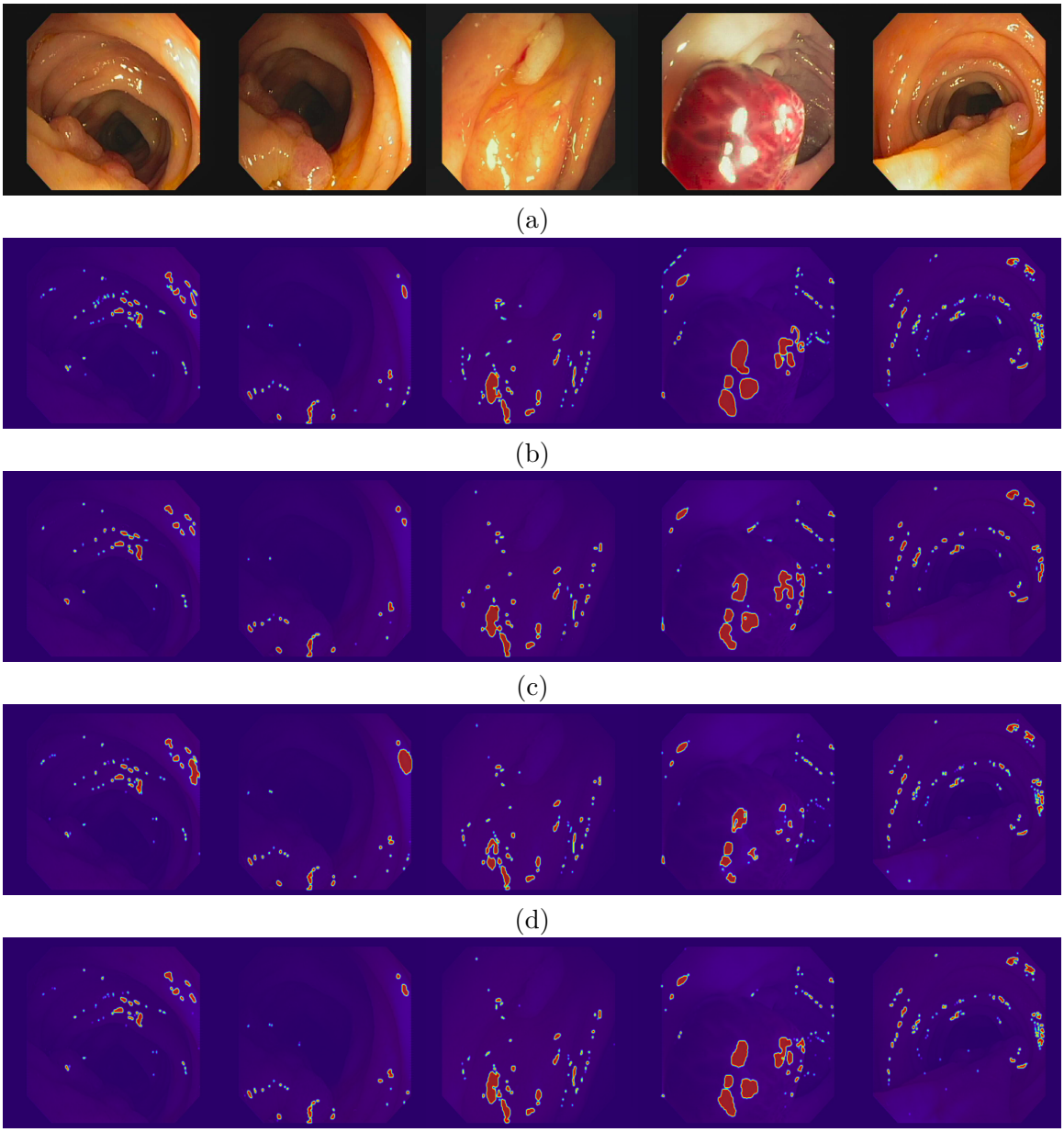


Figure 4.11: Heatmaps associated with segmentation results obtained from different deep learning models. (a) Original images; (b) Ground Truth; (c) DeepLabV3+[140]; (d) U-Net[66] and; (e) Our method

Analysis of Overexposed Regions

We considered some image patches affected by overexposure to study each model's potential in dealing with specular detection in such regions. We selected 19 images from the Test set and divided them into 6 equal-sized patches of 128×144. From the obtained patches, 23 patches are selected, which contain specular highlights within overexposed regions. The

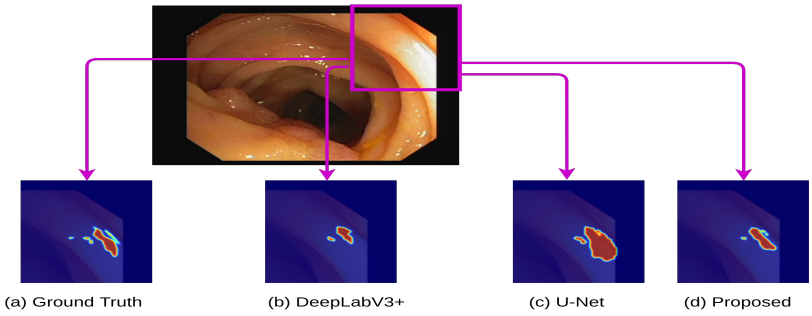


Figure 4.12: Performance of some deep learning models in overexposed regions

Table 4.4: Results in Overexposed regions

Method	**Precision (%)**	**Recall (%)**	**F1-score (%)**
DeeplabV3+ [140]	62.02	42.58	50.49
U-Net [66]	53.98	**76.54**	63.31
Proposed	**76.64**	72.31	**74.41**

segmented outputs of these patches are pixel-wise compared, and the corresponding precision, recall, and F1-score are noted. It can be observed from Table 4.4 that our model is more precise as compared to the other two models and hence reports fewer false positives. U-Net has achieved more recall rates, but its low precision rate indicates more false positives. DeepLabV3+ has attained the least recall value, but it has shown better performance in terms of precision and can be considered to be efficient in reducing false positives when compared to U-Net. The same can be seen in the heatmaps shown in Fig. 4.12.

Table 4.5: Histogram-based comparison with U-Net

Method	**Overall**		**Overexposed regions**	
	Chi-square	**Hellinger**	**Chi-square**	**Hellinger**
U-Net [66]	0.0036	0.0232	0.2026	0.0666
Proposed	**0.0019**	**0.0213**	**0.0280**	**0.0442**

Evaluating Statistical Significance

As U-Net performed far better than DeeplabV3+, we performed a statistical comparison between U-Net and the proposed model to understand the results better. We generated heatmaps, each for the ground truth and the segmented results obtained from our proposed model and U-Net. For each heatmap, a corresponding histogram is plotted, based on which

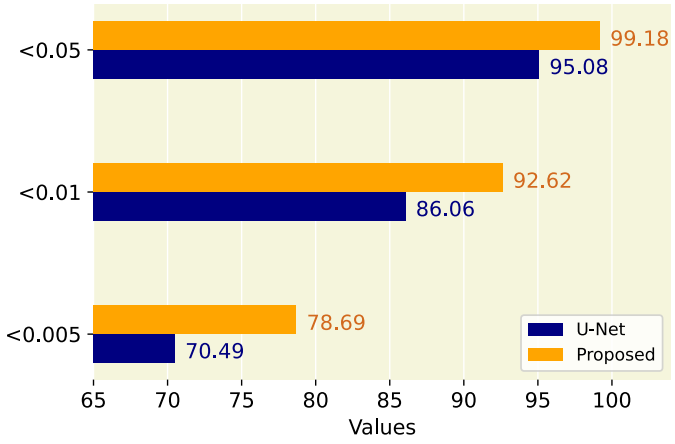


Figure 4.13: Comparative distribution of the number of images within the specified range of Chi-square values obtained from histogram comparison

statistical comparison is carried out. This involves matching the histogram of the ground truth heatmap with that of the U-Net's and the proposed model's corresponding heatmap. Chi-square (χ) distance and Hellinger (HG) distance are the two metrics adopted to express the matching of two histograms in terms of numerical values. Let h_1 and h_2 be the two histograms to be compared. χ and HG can be computed as:

$$\chi(h_1, h_2) = \sum_I \frac{(h_1(I) - h_2(I))^2}{h_1(I)} \tag{4.12}$$

$$HG(h_1, h_2) = \sqrt{1 - \frac{1}{\sqrt{\bar{h}_1 \bar{h}_2 \mathfrak{b}'^2}} \sum_I \sqrt{h_1(I) h_2(I)}} \tag{4.13}$$

where $\bar{h}_j = \frac{1}{\mathfrak{b}'} \sum_{J'} h_j(J')$ and $\mathfrak{b}'$ is the number of histogram bins. Low values of the specified metrics indicate a better match of the two histograms. Table 4.5 is comprised of related numerical values obtained after computing the average over all the test set images. The reported results illustrate the statistically better performance of our model over U-Net. To compare the image distribution with these outcomes, we divided the χ values into three ranges. Figure 4.13 shows that 78.69% of the images significantly match with their corresponding ground truths and attain a χ distance of less than 0.005. In the case of U-Net, the results are inferior to that of ours, with only 70.49% test images assigned within the range of 0.005. Further, the proposed model achieved better results, with 99.18% of the images reporting a χ distance of less than 0.05. Similarly, we compared histograms of different patches obtained during the analysis of the overexposed regions. The results

achieved by the proposed method are observed to be statistically more significant, with smaller χ as well as HG distance.

4.4 Chapter Summary

In this chapter, we proposed a GAN-based framework to translate uninformative colonoscopy frames into clinically significant frames. We showed that the translated frames improve polyp detection F1-score and mAP@0.5, with negligible reduction in mAP@0.5:0.95. We analyzed the types of artifacts where the CycleGAN performed well and identified the scope of improvements. Since the artifacts in colonoscopy video frames alter the various aspects of images, such as structure, texture, and color, this work lays the foundation for a more interesting future work of developing a standalone model to address all the artifacts in one go.

We also proposed a deep learning model to segment specular highlights in colonoscopy images. Our model integrates spectral information with CNNs and helps in extracting features at multi-resolution. The proposed method has adopted encoder-decoder architecture. Further, we have introduced DWTs and IDWTs instead of max-pooling and upsampling operations, respectively. This prevents loss of contextual and spatial information and assists in accurate segmented image reconstruction. We have modified the low-frequency sub-band obtained from DWT to lower the impact of overexposed regions. This improves specularity detection in those areas which suffer from overexposure due to illumination conditions. Our method has shown satisfactory results and has improved detection performance in overexposed regions, but there is a scope for improvement at the specular regions' boundaries. The boundaries could be more crisply segmented to reduce false positives. Analysis done in the overexposed regions indicates that there are some undetected highlights which have affected the recall rate.

5

Case Study and Dataset Design

Chapter Highlights

- The existing polyp segmentation (a task carried out after pre-processing of colonoscopy frames) literature lacks accurate algorithm comparisons due to inconsistent test data and closed source code.
- We demonstrate that for the reproducibility of different algorithms, we need an open-access dataset that covers a wide range of medical conditions.
- We present a case study that analyzes different algorithms developed on the same dataset and submitted to two challenges in the year 2020 and 2021.
- The case study highlights the advancements in polyp segmentation and encourages qualitative evaluation for building more transparent AI-based colonoscopy systems.
- We also curated and designed an open-access GI multi-class dataset acquired from two centers. It includes cases from pathological and normal findings, anatomical landmarks, and polyp removal.
- This chapter is based on the publications "*Validating Polyp and Instrument Segmentation Methods in Colonoscopy Through Medico 2020 and MedAI 2021 Challenges*", accepted for publication in Medical Image Analysis and "*GastroVision: A Multi-class Endoscopy Image Dataset for Computer Aided Gastrointestinal Disease Detection*", presented at ICML (ML4MHD) 2023.

This chapter studies the shortcomings of the existing polyp segmentation approaches and provides a detailed analysis of the related solutions. The segmentation is ideally performed over the keyframes after they are pre-processed for artifacts removal. Most existing polyp segmentation algorithms are developed using inconsistent dataset splits or private data, closed source or proprietary software, and methods that lack reproducibility. Therefore, to promote the development of efficient and transparent methods, we presented a case study on two challenges, namely, "Medico Automatic Polyp Segmentation (Medico 2020)" and "MedAI: Transparency in Medical Image Segmentation (MedAI 2021)" competitions. The Medico 2020 challenge received submissions from 17 teams, while the MedAI 2021 challenge also gathered submissions from another 17 distinct teams in the following year. We present a comprehensive summary and analyze each contribution, highlight the strength of the best-performing methods, and discuss the possibility of clinical translations of such methods. Our analysis revealed that the participants improved dice coefficient metrics from 0.8607 in 2020 to 0.8993 in 2021 despite adding diverse and challenging frames frequently missed during a routine clinical examination. The best team obtained a final transparency score of 21 out of 25. Moreover, this case study highlights the need for datasets that can cover a range of medical conditions and are easily accessible to the research community for better comparative analysis and reproducibility. However, the medical domain encounters a scarcity of large-scale, precisely labeled, and diverse datasets. This scarcity is attributed to the legal restrictions and extensive manual efforts required for accurate annotations from clinicians. To address these challenges, in this chapter, we further present GastroVision, a multi-class open-access Gastrointestinal (GI) endoscopy dataset that includes different anatomical landmarks, pathological abnormalities, polyp removal cases and normal findings (a total of 27 classes) from the GI tract. The dataset comprises 8,000 images acquired from Bærum Hospital in Norway and Karolinska University Hospital in Sweden and was annotated and verified by experienced GI endoscopists. Furthermore, we validate the significance of our dataset with extensive benchmarking based on the popular deep learning based baseline models. We believe our dataset can facilitate the develop-

ment of AI-based algorithms for GI disease detection and classification. Our dataset is available at https://osf.io/84e7f/.

5.1 Introduction

With the advent of deep learning, several solutions for polyp detection, segmentation, and classification have been proposed. These are the tasks that a CAD system performs to predict the diagnostic outcomes. The sequence of tasks undertaken by an ideal CAD system (discussed in Section 2.3) follows polyp detection, segmentation, and classification after performing keyframe selection and artifacts removal (discussed in the previous two chapters). Among these tasks, the most common and well-explored is the polyp segmentation. Despite numerous existing works, generating a comparative analysis of all these methods is difficult to achieve as they are not directly comparable. Some of the reasons for such an issue include inconsistent dataset split, use of private datasets, and closed sources, making reproducibility difficult. Moreover, most methods lack the interpretability and transparency required to understand the predictions made by the model. Therefore, we present a comprehensive analysis of the results of the two prominent challenges in the field of automatic polyp segmentation, namely, *"Medico Automatic Polyp Segmentation (Medico 2020)"* [1] challenge and the *"MedAI: Transparency in Medical Image Segmentation (MedAI 2021)"* [2] challenge. These challenges were designed to foster the development of CAD solutions on the same datasets, with a focus on transparency, explainability, robustness, speed, and generalization, aiming to evaluate the relevance of such algorithms in clinical workflows. The challenges provided posed four distinct tasks:

- Accurate polyp segmentation task to develop state-of-the-art algorithms for early detection and treatment of colon cancer (Medico 2020, MedAI 2021).
- Algorithm efficiency task to develop methods with the least Frames Per Second (FPS) on predetermined hardware (Medico 2020).
- Surgical instrument segmentation task to enable tracking and localization of essential tools in endoscopy and help to improve targeted biopsies and surgeries in complex GI tract organs (MedAI 2021).

[1] https://multimediaeval.github.io/editions/2020/tasks/medico/
[2] https://www.nora.ai/competition/image-segmentation.html

- Transparency task to evaluate the proposed system from a transparency point of view (for example, explanations of the training procedure, amount of data used and model's predictions interpretation) (MedAI 2021).

There are several other challenges conducted in the past such as GIANA 2017 [143], GIANA 2018 [122, 123], EndoCV 2021 [94, 144], and Medico 2021 [145]. These challenges motivate to have such platforms to develop state-of-the-art algorithms for critical medical issues. However, dataset availability is an important aspect of successfully organising these challenges. To design a system with accurate comparison and to provide the research community with the possibility of reproducibility, it is necessary to make medical data easily accessible. This allows researchers to contribute in the area and assess their work with respect to state-of-the-art techniques on comparable grounds. Although our case study is based on a single class (particularly polyp) and provides opportunity for reproducibility and transparency, it would be more beneficial from both research and clinical perspective if we could extend this practice to multi-class problems. As already discussed in Section 2.3.6, the currently available endoscopy datasets mainly focus on a single lesion, i.e., polyps. Moreover, some of these datasets are available on request and cover a few pathological classes. Therefore, our case study is followed by curating and designing of a multi-class GI dataset. Some more facts supporting the purpose of developing this dataset are discussed below.

- Most existing AI models depend on data acquired from a single center, which makes them less valid when faced with a varied patient population. This leads to spectrum bias under which AI systems encounter performance drops due to the significant shift in the original clinical context and the test sample population. In such cases, unexpected outcomes and diagnostic accuracy could be obtained using automated tools. Such bias issues could reach the clinical systems at any point of the process, including data collection, study design, data entry and pre-processing, algorithm design, and implementation. The very beginning of the process, i.e., data collection, is of utmost importance for reproducibility and to perform validations on images from a diverse population, different centers, and imaging modalities.

- To develop scalable healthcare systems, it is vital to consider performing real-time validations. However, the scarcity of comprehensive data covering a range of real-time imaging scenarios arising during endoscopy or colonoscopy makes it difficult to develop a robust AI-based model. Although much progress has been made on

automated cancer detection and classification [17, 62], it is still challenging to adapt such models into real-time clinical settings as they are tested on small-sized datasets with limited classes.

Some classes in the dataset could be scarce because some conditions or diseases occur less often. Consequently, such findings are not frequently captured and remain unexplored despite requiring medical attention. AI-based detection of these findings, even with a small sample count, can significantly benefit from techniques like one-shot or few-shot learning. These techniques allow the AI models to learn patterns and features indicative of the condition, thus, enabling accurate diagnosis with minimal training data. Therefore, in this chapter, we publish *GastroVision*, an open-access multi-class endoscopy image dataset for automated GI disease detection that does not require prior consenting and can be downloaded easily with a single click. The data covers a wide range of classes that can allow initial exploration of many anatomical landmarks and pathological findings.

5.1.1 Contributions

The main contributions of this chapter are summarized below:

- **Case Study**
  - We present a comprehensive and detailed analysis of all participant results.
  - We provide an overview and comparative analysis of the developed methods.
  - We obtain and discuss new insights into the current state of AI in the field of GI endoscopy including open challenges and future directions.
  - We provide a detailed discussion of issues such as generalizability issues, multi-center and out-of-distribution testing in context to current limitations of CAD systems.
- **Dataset Curation and Design**
  - We present an open-access multi-class GI endoscopy dataset, namely, *Gastro-vision*, containing 8,000 images with 27 classes from two hospitals in Norway and Sweden. The dataset exhibits a diverse range of classes, including anatomical landmarks, pathological findings, polyp removal cases and normal or regular findings. It covers a wide range of clinical scenarios encountered in endoscopic procedures.

– We evaluated a series of deep learning baseline models on standard evaluation metrics using our proposed dataset. With this baseline, we invite the research community to improve our results and develop novel GI endoscopy solutions on our comprehensive set of GI finding classes. Additionally, we encourage computer vision and machine learning researchers to validate their methods on our open-access data for a fair comparison. This can aid in developing state-of-the-art solutions and automated systems for GI disease detection and other general machine learning classification tasks.

These contributions are explained in detail in the subsequent sections. Section 5.2 delves into the case study, while Section 5.3 provides a detailed description of the dataset.

5.2 Case Study

5.2.1 Challenge Description

Medico 2020 Automatic Polyp Segmentation Challenge

Medico 2020 Automatic Polyp Segmentation Challenge, hosted through MediaEval platform (Multimedia Evaluation Workshop), is aimed at benchmarking automated polyp segmentation algorithms. The participating teams involving researchers from medical image analysis, multimedia, machine learning, and computer vision were provided with the same dataset. Participants could use any method, focusing on creating automated solutions. The submissions were open for two tasks, namely, automatic polyp segmentation and algorithmic efficiency tasks.

Automatic Polyp Segmentation Task: This task was based on developing innovative and efficient polyp segmentation algorithms which at the same time, were expected to be fast enough for clinical deployments. To participate in the challenge, participants were required to train their segmentation models on an available training set. Once the test set was released, participants could test their models and submit their predicted segmentation maps to the organizers in a .zip file with the name of each segmentation map image matching the colonoscopy image in the test set.

Algorithmic Efficiency Task: CAD systems deployed in clinical settings need to operate in real-time; however, such systems often have fewer parameters and lower accuracy than computationally intensive algorithms. To address this trade-off, we encouraged the teams to propose a solution that is lightweight but also prioritizes segmentation accuracy.

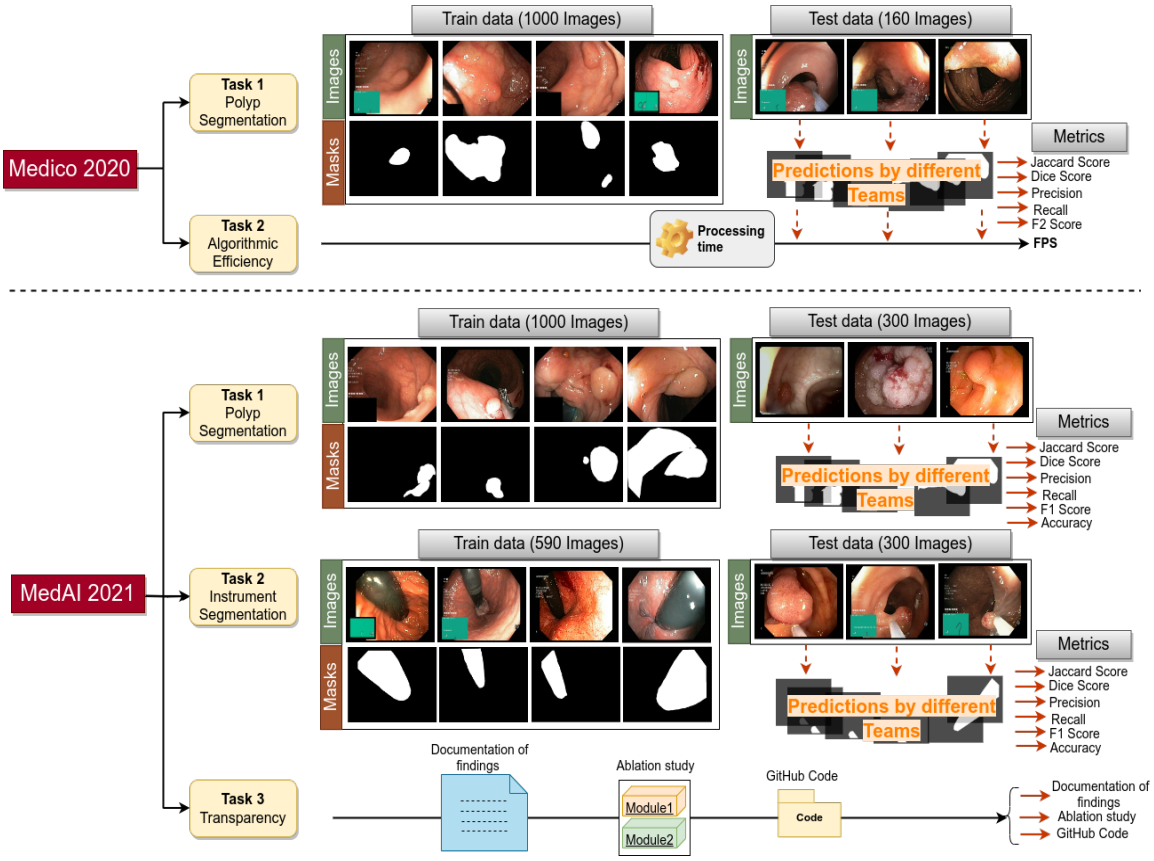


Figure 5.1: The overview of the "***Medico 2020 Polyp***" and "***MedAI 2021 Transparency*** " challenges. We describe each task along with the number of training and testing datasets and the evaluation metrics used in the tasks.

Participants were asked to submit docker images of their proposed algorithms. These algorithms were then evaluated on a dedicated Nvidia GeForce GTX1080 graphics card. A threshold for mIoU was set for the validity of an algorithm's efficiency. After this validation check, the teams were ranked based on the FPS.

MedAI: Transparency in Medical Image Segmentation Challenge

MedAI: Transparency in Medical Image Segmentation Challenge was held at the Nordic AI Meet 2021 (Nordic Young Researchers Symposium). It focused on segmentation tasks and transparency in machine learning based solutions. The two segmentation tasks included polyp segmentation and instrument segmentation. Similar to the other challenge, participants were granted the flexibility to use any method, focusing on developing automated solutions.

Automatic Polyp Segmentation Task: The main objective of this task was similar

to that of the first task of the previous competition. The participants were asked to submit a .zip file with predicted binary masks in a .png format. The test data provided to the participants was without a ground truth. Hence, with such hidden test data, the results were reliable and provided a valuable benchmark for the field.

Automatic Instrument Segmentation Task: This task also focused on segmentation, but the ROIs considered here are the GI accessory instruments such as biopsy forceps or polyp snares used during live endoscopy procedures. Performing segmentation of these instruments helps with the tracking and localization of essential tools in endoscopy that could aid endoscopists during interventions (such as polypectomies). The submission and evaluation procedure was the same as that of the previous task.

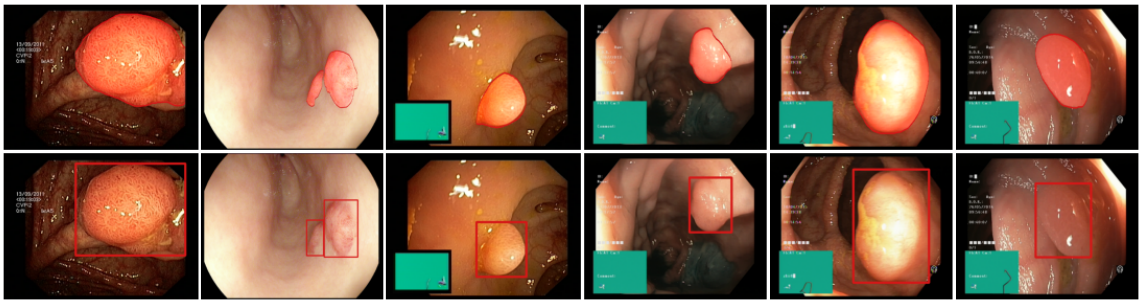
Transparency Task: This task is targeted to assess the submissions from a transparency perspective. It included evaluation on the basis of the interpretation of the model's predictions and the use of explainable AI methods. To achieve this goal, the participants were encouraged to perform ablation studies and share their code on the GitHub repository with detailed procedures on how to reproduce the results. Additionally, the teams were asked to provide a one-page document with explanations related to their transparency task outcomes.

Figure 5.1 provides an overview of both challenges along with the total number of images used for training and testing in each task. Ground truth samples with their corresponding original images are also presented for the segmentation tasks. In addition, task-specific metrics are presented (for example, FPS for "Algorithm efficiency").

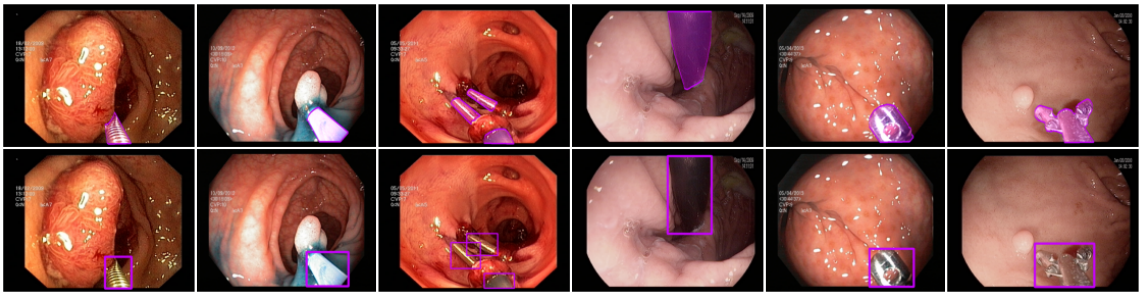
5.2.2 Challenge Datasets and Methods

Challenge Datasets

The datasets contain 1,000 polyp images and their corresponding ground truth mask taken from Kvasir-SEG [9]. The datasets were acquired from real routine clinical examinations at Vestre Viken Health Trust (VV) in Norway by a team of expert gastroenterologists. The resolution of images varies from 332×487 to 1920×1072 pixels. Some images contain green thumbnails in the lower-left corner of the images showing the position marking from the ScopeGuide (Olympus). To extend the dataset to the segmentation class, a team of experienced engineers and expert gastroenterologists annotated the polyp images using the label box tool. Once the ground truth was created, the images and ground truths were combined to facilitate the review process. These images were sent to a team of expert gastroenterologists for validation through a web-based interface. The data proportion for



(a) Examples samples from the test data of Medico 2020 (first three columns) and MedAI 2021 (last three columns) for the polyp segmentation task.



(b) Example samples from the MedAI 2021 Instrument segmentation task.

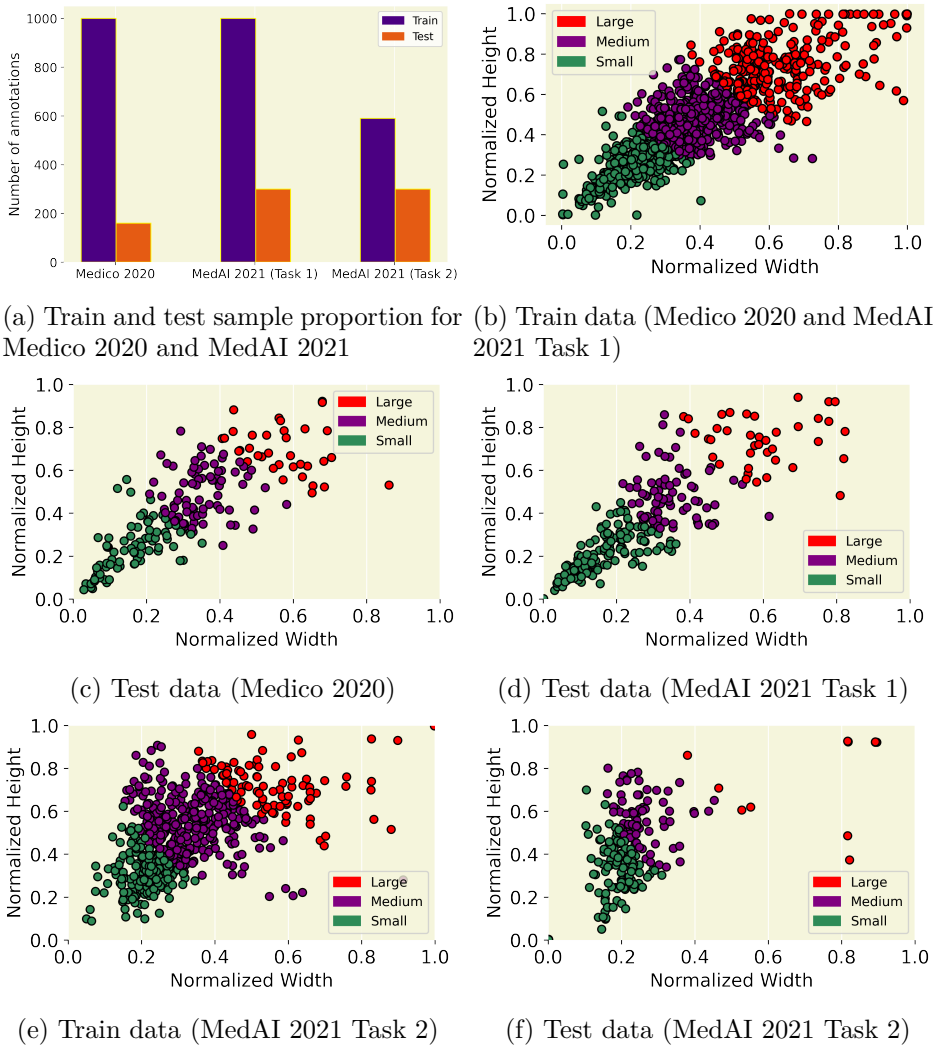
Figure 5.2: Example of the test datasets from the Medico 2020 and MedAI 2021 datasets.

each training and test set followed the general split ratio used in the literature. The training dataset has been made publicly available as open access and is widely available[3]. This dataset was used as a training set in both challenges. The other dataset details pertaining to individual challenges are given below:

Medico 2020: The test dataset contains unique polyp images encompassing a wide range of diverse clinical scenarios with different polyp characteristics, varying lighting conditions and image resolution, low-quality images, as well as complex polyp images (for example, with instruments and residual stool) that the model has never encountered before. Only the organizers had access to the test case labels. Currently, the test data can be downloaded from[4]. Some samples are shown in Figure 5.2a.

MedAI 2021: The Kvasir-SEG [9] dataset was used as the development set for the polyp segmentation task. Similarly, Kvasir-Instrument [146] was used as the training dataset for the instrument segmentation task. It can be downloaded from[5]. Some sample images for polyp segmentation and instrument segmentation tasks are presented in Figure 5.2a and Figure 5.2b.

[3] https://datasets.simula.no/kvasir-seg/
[4] https://drive.google.com/file/d/1uP2W2g0iCCS3T6Cf7TPmNdSX4gayOrv2
[5] https://datasets.simula.no/kvasir-instrument/



(a) Train and test sample proportion for Medico 2020 and MedAI 2021

(b) Train data (Medico 2020 and MedAI 2021 Task 1)

(c) Test data (Medico 2020)

(d) Test data (MedAI 2021 Task 1)

(e) Train data (MedAI 2021 Task 2)

(f) Test data (MedAI 2021 Task 2)

Figure 5.3: Data distribution details of train and test sets used in Medico 2020 and MedAI 2021 challenges. Large, medium, and small represent the distribution information of regions of interest in the data samples.

Figure 5.3 shows the data distribution of the train and test datasets used in Medico 2020 and MedAI 2021. We have categorized the images into "small", "medium" and "large" according to the size of ROI using a randomly selected threshold of 0.3 and 0.1 and plotted the normalized height versus normalized width of each data point. This is to visualize the dimension of each data point and observe the diversity and complexity of the dataset used in the study. The information about the size categories and the dataset's dimensions is crucial for assessing the performance and robustness of the proposed algorithms.

Methods

The Medico 2020 challenge received submissions from 17 teams, while the MedAI 2021 challenge also gathered submissions from 17 teams in the following year. Table 5.1 and Table 5.2 present details of algorithms developed and used by different teams. These details include the overall method, the backbone architecture, the loss function, and the optimizer adopted by the individual team. In the Medico 2020 challenge, all teams participated in Task 1, whereas only 9 teams provided submissions for Task 2. In the MedAI 2021 challenge, most of the teams participated in all three tasks except for four, which participated in either one or two sub-tasks. Most participating teams have used the same architecture in their submission for both subtasks. However, two teams, namely *Vyobotics* [147] and *MedSeg_JU* [148] have participated in only one of the subtasks. The team *Vyobotics* [147] has participated in the polyp segmentation task whereas the team *MedSeg_JU* [148] has participated in the accessory instrument segmentation task.

5.2.3 Results

In this section, we present a summary of the evaluated results obtained on the test dataset by all the participating teams in the two challenges: "Medico 2020" and "MedAI 2021". Each challenge consists of tasks with a specific focus and evaluation metrics. There were two tasks for the Medico 2020 challenge, namely *polyp segmentation* and *algorithm efficiency* tasks. In the MedAI 2021, there were three tasks, namely *polyp segmentation*, *endoscopic accessory instrument segmentation* and *transparency task*. The teams were evaluated based on standard evaluation metrics such as mIoU, DSC, Rec, Pre, Acc, F1, F2, and FPS. We emphasized mIoU, DSC, and FPS more, whereas we also acknowledged the importance of recall and precision as they are useful metrics in clinical settings.

Medico 2020 Results

Polyp Segmentation Task

In Table 5.3, we provide the results for the *polyp segmentation* task. It can be observed that Team "PRML2020GU" outperforms other participating teams in the polyp segmentation task. It achieves a mIoU of 0.7897, DSC of 0.8607, recall of 0.9031, precision of 0.8673, and F2 of 0.8748. Team "HBKU_UNITN_SIMULA" was the second best-performing team with mIoU of 0.7773. similarly, "AI-TCE" was the third best-performing team with mIoU of 0.7773. The best-performing team, "PRML2020GU," used an encoder-decoder struc-

Table 5.1: Summary of the participating teams algorithm for Medico 2020.

Team Name	Algorithm	Backbone	Nature	Choice basis	Aug.	Loss	Optimizer
FAST-NU-DS [149]	Depth-wise separable convolution and ASPP	ResUNet++	Cascade of depth-wise separable convolutions	mIoU and DSC	Yes	IoU	Adam
AI-TCE [150]	Multi-Supervision Net	EfficientNetB4	Encoder-Multi Supervision Decoder	Acc and DSC	Yes	Categorical cross-entropy + DSC loss	Adam
ML-MMIV SARUAR [151]	Encoder-decoder based architecture based on ResNet50	ResNet50	Cascade of residual blocks	mIoU and DSC	Yes	cross-entropy	Adam
UiO-Zero [152]	GAN	None	GAN with CNN based generator and discriminator	Image-to-image translation	No	Standard conditional GAN adversarial loss	Adam
HBKU UNITN SIMULA [153]	Residual module, Inception module, Adaptive CNN with U-Net and PraNet	U-Net and Resnet50	Cascade of residual blocks and inception module	mIoU and DSC	Yes	Bce + Dsc loss	Adam
AI-JMU [154]	Cascade Mask R-CNN	ResNeSt backbone, Cascade Architecture	Deep CNN	DSC and mIoU	Yes	Binary cross-entropy	SGD
SSB [155]	U-Net	ResNet-34, EfficientNet-B2	Ensemble	DSC and mIoU	Yes	Tversky loss	Adam
AMI LAB [156]	Knowledge distillation on ResUNet++	ResUNet++	Ensemble	mIoU and DSC	Yes	Distillation loss	Adam
UNITRK [157]	Knowledge transfer using UNet	Pre-trained U-Net model	Encoder-decoder	mIoU and DSC	Yes	Compound loss of DSC and BCE	Adam
MedSeg_JU [158]	Conditional GAN (cGAN)	None	Encoder-decoder	mIoU and DSC	Yes	Weighted loss of MSE and BCE	Adam
IIAI-Med [159]	PraNet	Res2Net	Encoder-decoder	mIoU, DSC and FPS	No	weighted IoU loss + BCE loss	Adam
HGV-HCMUS [160]	PraNet and ResUNet++ with triple path	ResUNet++	Encoder-decoder	mIoU	Yes	Categorical crossentropy	Adam
GS-CDT [161]	U-Net	None	Encoder-decoder	Acc and Speed	Yes	Non-Binarized mIoU	Adam
PRML20202GU [162]	Efficient-UNet +Channel-Spatial Attention + Deep Supervision	Variants of EfficientNet	Encoder-decoder	mIoU and DSC	Yes	BCE + DSC loss	Adam
VT [163]	U-Net coupled with PYRA	None	Encoder-decoder	mIoU and DSC	Yes	BCEWithLogits loss	RMSprop
IRISNSYSU [164]	Temporal-Spatial Attention Model	Faster-RCNN	Hybrid attention interface	AP	Yes	Cross entropy	Adam
NTK [165]	Residual blocks combined with SE network	None	Encoder-decoder	DSC, mIoU and FPS	No	BCE + DSC loss	Adam

Table 5.2: Summary of the participating teams algorithm for MedAI 2021.

Team Name	**Segmentation Task**	**Algorithm**	**Backbone**	**Nature**	**Choice basis**	**Augmentation**	**Loss**	**Optimizer**
The Segmentors [166]	Polyp, Instrument	U-Net	None	Encoder-decoder	DSC and mIoU	Yes	DSC	Adam
The Arctic [167]	Polyp, Instrument	DeeplabV3plus + ResNet101	None	Hybrid	DSC	Yes	Cross-entropy	Adam
mTEC [168]	Polyp, Instrument	DPRA-EdgeNet	HarDNet	Cascade	DSC and mIoU	No	(Dice + BCE) loss	Adam
MedSeg_JU [148]	Instrument	EM-Net	EfficientNet-B3	Encoder-decoder	DSC	Yes	DSC	Adam
MAHUNM [169]	Polyp, Instrument	DeeplabV3 with GRU	ResNet-50/ResNet-101	Sequential	DSC and mIoU	No	BCE With Logits Loss	Adam
IIAI-CV&Med [170]	Polyp, Instrument	Polyp-PVT, Sinv2-PVT and Transfuse- PVT	Transformer	Ensemble	Majority voting	No	IoU	Adam
NYCity [171]	Polyp, Instrument	HarDNet-85, ResNet-101	Transformer	Ensemble	Accuracy	Yes	IoU	Gradient centralization
PRML [172]	Polyp, Instrument	Ef-UNet	EfficientNet	Encoder-decoder	DSC and mIoU	No	DSC Loss	Adam
leen [173]	Polyp, Instrument	GAN	None	Encoder-decoder	DSC and mIoU	No	BCE and L1 loss	Adam
CV&Med IIAI [174]	Polyp, Instrument	SINetv2	PVT v2	Encoder-decoder	mIoU	No	PPA loss	Adam
Polypixel [175]	Polyp, Instrument	Transfer learning using EfficientNet B1	None	CNN	DSC and mIoU	Yes	IoU	Adam
agaldran [176]	Polyp, Instrument	Double Encoder-Decoder with temperature scaling Feature	Pyramid Network as Decoder and Resnext101 as pretrained decoder	Sequential	DSC	Yes	DSC	Sharpness-aware minimization(SAM)+ Adam
TeamAIKitchen [177]	Polyp, Instrument	U-Net	None	Encoder-decoder	DSC	Yes	DSC	Adam
CamAI [178]	Polyp, Instrument	Transfer learning (Attention U-Net)	ResNet-152	Ensemble	Accuracy	Yes	Unified focal loss	SGD
OXGastroVision [179]	Polyp, Instrument	DDANet + FANet	None	Encoder-decoder	DSC	No	BCE and DSC loss	Adam
Vyobotics [147]	Polyp	DDANet	None	Encoder-decoder	DSC and mIoU	Yes	BCE and DSC loss	Adam
NAAMII [147]	Polyp, Instrument	U2Net	None	Encoder-decoder	mIoU	Yes	Mean Squared Error, Cross-entropy	Adam

Table 5.3: Performance comparison on Polyp segmentation task (Medico 2020). 'Bold' refers to the best score, and 'underline' refers to the second-best score. We follow this consistently in all the Tables. ↑ indicates a higher value is better.

Team Name	**mIoU ↑**	**DSC ↑**	**Recall ↑**	**Precision ↑**	**F2 ↑**
PRML2020GU	**0.78975**	**0.86076**	0.90312	0.86731	0.87481
HBKU_UNITN_ SIMULA	0.77736	0.84768	0.85034	0.88971	0.84483
AI-TCE	0.77733	0.85030	**0.91646**	0.83897	**0.87901**
HGV-HCMUS	0.76597	0.84050	0.89439	0.84455	0.85768
IIAI-Med	0.76195	0.83854	0.83049	**0.90121**	0.82837
SBS	0.75503	0.83162	0.83168	0.88513	0.82490
ML-MMIV Saruar	0.75168	0.82289	0.83908	0.88228	0.82492
AI-JMU	0.73742	0.81437	0.82661	0.87432	0.81038
MedSeg_JU	0.71330	0.80195	0.83542	0.82864	0.81240
VT	0.70578	0.79264	0.88353	0.78784	0.82368
NKT	0.68473	0.78012	0.80771	0.81264	0.78546
UNITRK	0.64379	0.72878	0.70989	0.85726	0.71312
GeorgeBatch	0.63511	0.73276	0.75003	0.82294	0.73615
AMI Lab	0.61958	0.70889	0.72865	0.79140	0.71226
IRIS-NSYSU	0.50353	0.64173	0.87915	0.58498	0.75089
UiO-Zero	0.43814	0.56185	0.69721	0.55587	0.61102
FAST-NU-DS	0.18344	0.26691	0.27447	0.29184	0.26762

ture with EfficientNet as the backbone and a U-Net decoder with channel-spatial attention and deep supervision. This architecture had an improvement of 1.23% and 1.30% over the mIoU and DSC achieved by the Team "HBKU_UNITN_SIMULA", which used an average of three PraNet and five ResUNet++ trained on different training and validation datasets.

Table 5.4: Algorithm efficiency task for polyp segmentation (Medico 2020). Note that some teams provided the same solution for this task as used in Task 1, whereas others designed different architecture specifically for the efficiency task (Task 2). ↑ indicates a higher value is better.

Team Name	**mIoU ↑**	**DSC ↑**	**Recall ↑**	**Precision ↑**	**F2 ↑**	**FPS ↑**
HCMUS	**0.7364**	0.8074	0.8164	**0.8646**	0.8067	33.27
SBS	0.7341	**0.8148**	0.8764	0.8145	**0.8354**	26.66
NKT	0.6847	0.7801	0.8077	0.8126	0.7854	80.60
FAST-NU-DS	0.6582	0.7556	**0.8982**	0.7171	0.8109	67.51
UNITRK	0.6437	0.7287	0.7098	0.8572	0.7131	116.79
GeorgeBatch	0.6351	0.7327	0.7500	0.8229	0.7361	**196.79**
AMI Lab	0.6195	0.7088	0.7286	0.7914	0.7122	107.87
AI-JMU	0.7213	0.8017	0.8359	0.8495	0.8056	3.36
PRML2020GU	0.5083	0.6265	0.6003	0.7870	0.6029	2.25

Algorithm Efficiency Task

For the second task, as in Table 5.4, team "PRML2020GU" has poor speed performance with a processing speed of only 2.25 fps, which is not desirable for a real-time efficient model. An interesting observation is that Team "GeorgeBatch" outperforms other participating teams in the algorithm efficiency task with a processing speed of 196.79 fps, as seen from Table 5.4. However, it is worth noting that the team obtained a low mIoU of 0.6351 for the polyp segmentation task, even though we are considering it as the winner in this task. Team "UNITRK" obtained a second-best fps of 116.79. Similarly, team "NKT" obtained a balanced mIoU of 0.6847 and a high speed of 80.60 fps, and was ranked third for this task. Despite the two teams, "UNITRK" and "GeorgeBatch", achieving the highest evaluation fps values, there is a trade-off between speed and mIoU. Low FPS cannot be used for real-time medical processing applications, and low overlap evaluation metrics cannot generate precise segmentation masks.

MedAI 2021 Challenge Results

Polyp Segmentation Task

In Tables 5.5, we tabulated the evaluation results of all the participating teams in MedAI 2021 for polyp segmentation task. The table shows that team "agaldran" outperforms other teams in the polyp segmentation task with mIoU of 0.8522 and DSC of 0.8965. Team "CV&Med IIAI" also showed good performance and was ranked 2^{nd} in the polyp segmentation task with a mIoU of 0.8484, a very small difference from the best-performing team.

Instrument Segmentation Task

From Table 5.6, it can be observed that the same team, "agaldran", also outperforms other participating teams in the instrument segmentation task with a high mIoU of 0.9364 and DSC of 0.9635. Team "NYCity" was ranked 2^{nd} in this task with a mIoU of 0.9326 and DSC of 0.9586. However, Team "NYCity" obtained the highest recall of 0.9712, which signifies it has low false negative (FN) regions in the predicted segmentation mask compared to team "agaldran". Another interesting observation is the team "agaldran" also achieved higher metric values for the instrument segmentation task as compared to the polyp segmentation task, as instrument segmentation is relatively easier than polyp extraction due to the greater variability of the latter regarding color and appearance.

Transparency Task

We present the transparency results in Table 5.7. Team "agaldran" outperformed

Table 5.5: Performance evaluation for the participating teams for the polyp segmentation task in MedAI 2021 Challenge. ↑ indicates a higher value is better.

Team Name	**mIoU ↑**	**DSC ↑**	**Recall ↑**	**Precision ↑**
agaldran	**0.8522**	**0.8965**	0.9009	0.9242
NYCity	0.8418	0.8885	0.8794	**0.9319**
IIAI-CV&Med	0.8361	0.8927	**0.9195**	0.8963
mTEC	0.8334	0.8892	0.9010	0.9096
CV&Med IIAI	0.8213	0.8612	0.8602	0.8814
PRML	0.8116	0.8669	0.8852	0.8922
CamAI	0.8083	0.8701	0.8702	0.9052
The Arctic	0.8022	0.8533	0.8604	0.8821
Polypixel	0.7997	0.8567	0.8868	0.8659
MAHUNM	0.7495	0.8189	0.8397	0.8568
OXGastroVision	0.7334	0.7966	0.8158	0.8374
Vyobotics	0.7220	0.7967	0.8214	0.8359
NAAMII	0.6041	0.6940	0.7499	0.7334
leen	0.4595	0.5531	0.6389	0.5860
The Segmentors	0.3789	0.4205	0.4178	0.4640
TeamAIKitchen	0.2904	0.4100	0.7152	0.4910

Table 5.6: Performance of participating teams for instrument segmentation task of MedAI 2021 Challenge. ↑ indicates a higher value is better.

TeamName	**mIoU ↑**	**DSC ↑**	**Recall ↑**	**Precision ↑**
agaldran	**0.9364**	**0.9635**	0.9692	**0.9632**
NYCity	0.9326	0.9586	**0.9712**	0.9516
mTEC	0.9245	0.9553	0.9687	0.9490
PRML	0.9178	0.9528	0.9687	0.9441
IIAI-CV&Med	0.9148	0.9490	0.9612	0.9473
CV&Med IIAI	0.9136	0.9512	0.9605	0.9500
Polypixel	0.9114	0.9478	0.9591	0.9438
CamAI	0.9079	0.9442	0.9527	0.9468
The Arctic	0.9078	0.9448	0.9735	0.9231
OXGastroVision	0.8692	0.9073	0.9236	0.9096
MAHUNM	0.8523	0.9080	0.9535	0.8864
MedSeg_JU	0.8205	0.8632	0.9005	0.8464
TeamAIKitchen	0.7257	0.7980	0.7955	0.8510
leen	0.6991	0.7845	0.7963	0.8232
NAAMII	0.6857	0.7741	0.8321	0.7669
The Segmentors	0.3668	0.3971	0.3985	0.4040

other competitors with a final score of 21 out of 25. Similarly, "mTEC" obtained a score of 17 out of 25 and was ranked 2^{nd}. Likewise, team "CamAI" obtained a score of 16 out of 25 and was ranked third in the transparency task. There were also efforts from

teams such as "The Arctic", which obtained a score of 13, and "IIAI-CV&Med", which obtained a score of 10. These scores show their effort to provide a transparent solution to the polyp and instrument segmentation tasks. We provide the final ranking and task-wise scores in Figure 5.4. Notably, team ***"agaldran"*** outperformed others in all three tasks and overall challenge and emerged as the winner of the MedAI 2021 challenge. Overall, ***"mTec"*** secured the second position. Following closely behind, ***"CamAI"*** showcased the third-best solution. The overall rank was computed by combining the mIoU scores of polyp and instrument segmentation tasks with the transparency score.

Table 5.7: Evaluation of the **'Transparency tasks'** for MedAI 2021 Challenge. For this task, a team of experts accessed the submission based on several criteria and provided a score based on the availability and quality of the source code (for e.g., open access, public availability, and documentation for reproducibility), model evaluation (for e.g., failure analysis, ablation study, explainability, and metrics used) and qualitative evaluation from clinical experts (e.g., usefulness and understandability of the results). Here, '0' refers to no submissions. Doctor evaluation was only calculated for the team whose manuscripts were accepted.

Team Name	Open Source			Model Evaluation				Doctor Evaluation		Final Score
	Publicly available (0 or 1)	Code Quality (0-3)	Readme (0-3)	Failure Analysis (0-3)	Ablation Study (0-3)	Explainability (0-3)	Metrics Used (0 or 1)	Usefulness (0-3)	Understandable (0-5)	
agaldran	1	2	3	3	3	3	1	2	3	21
CamAI	1	1	1	2	1	2	1	2	5	16
CV&Med IIAI	0	1	0	1	0	0	1			3
IIAI-CV&Med	1	1	2	0	0	0	1	1	4	10
leen	0	1	0	0	0	2	1			4
MAHUNM	1	1	0	0	0	0	1			3
mTEC	1	1	3	3	1	0	1	3	4	17
NYCity	0	0	0	0	0	0	1			1
OXGastroVision	0	2	0	0	0	0	1			3
Polypixel	1	1	2	0	0	0	1			5
PRML	0	1	0	0	0	0	1			2
TeamAIKitchen	0	1	0	0	0	0	1			2
The Arctic	1	2	1	1	0	3	1	1	3	13
The Segmentors	0	0	0	0	0	0	1			1

Figure 5.5a illustrates the plot of mIoU reported by each team in their submissions in the two challenges with three different tasks. It can be observed that the *polyp segmentation task* from 2020 to 2021 gained improvement with a larger number of submissions, achieving a mIoU of more than 0.80 and the best-performing team with a mIoU of around 0.85. Similar progress can be observed in Figure 5.5b where an overall mIoU increased by 4.93% when an average score is computed over all participating teams' individual best mIoU in the 2021 polyp segmentation challenge. We further compared all segmentation metrics, including DSC, recall, precision, mIoU score, accuracy, and F2 score, as shown in Figure 5.5c. Notably, the different evaluation metrics scores are consistent with instrument segmentation tasks in the MedAI challenge. However, there is a high variation in the mIoU between the different teams in the polyp segmentation tasks of Medico 2020 and MedAI 2021 challenges.

These values pertain to the best score corresponding to a particular metric the indi-

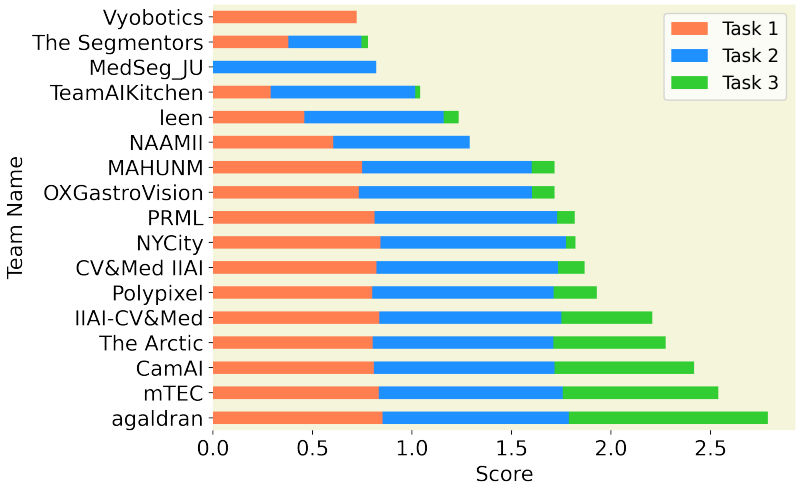
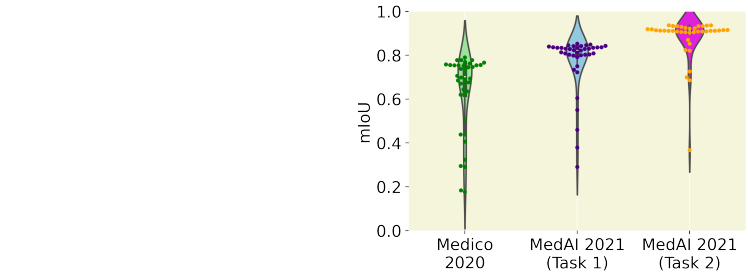


Figure 5.4: Task-wise scores achieved by participating teams of MedAI 2021 challenge. Team rankings are decided on the basis of overall scores in all three tasks.

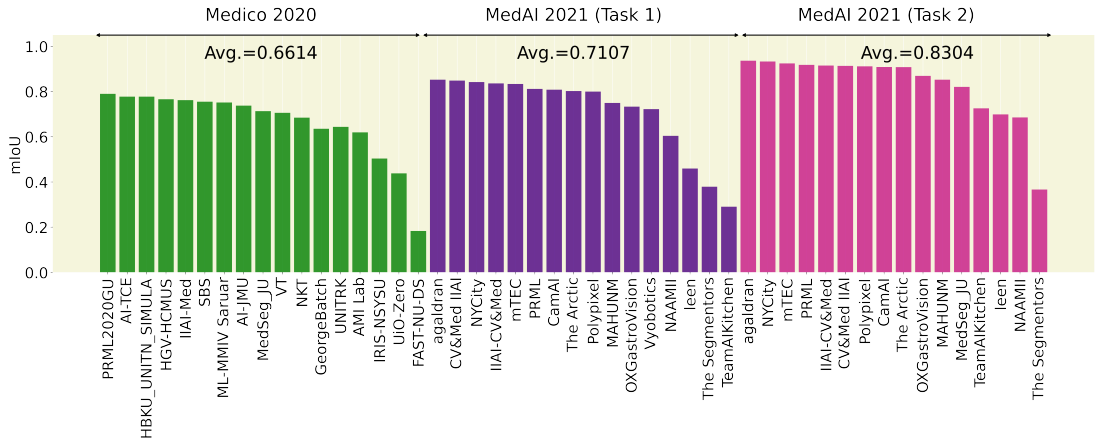
vidual team obtained in different executions. It is to be noted that each team was given the opportunity to submit five different submissions, and the best results for the best submission are reported in the Tables here. From here, it can be observed that most teams in the MedAI 2021 challenge reported overall high scores in terms of various segmentation metrics when compared to Medico 2020 outcomes, thus highlighting the improved performance trends in automated systems over time. Furthermore, it can also be visualized that unlike the high variations shown by teams' scores in the polyp segmentation task, better performance and smaller deviations in scores are reported in the instrument segmentation task. The high variations in the polyp segmentation results also show that polyp segmentation is more challenging because of the presence of variations in the size, structure and appearance of the polyps, and the presence of the artifacts and lighting conditions deteriorate the algorithm's performance.

5.2.4 Discussions

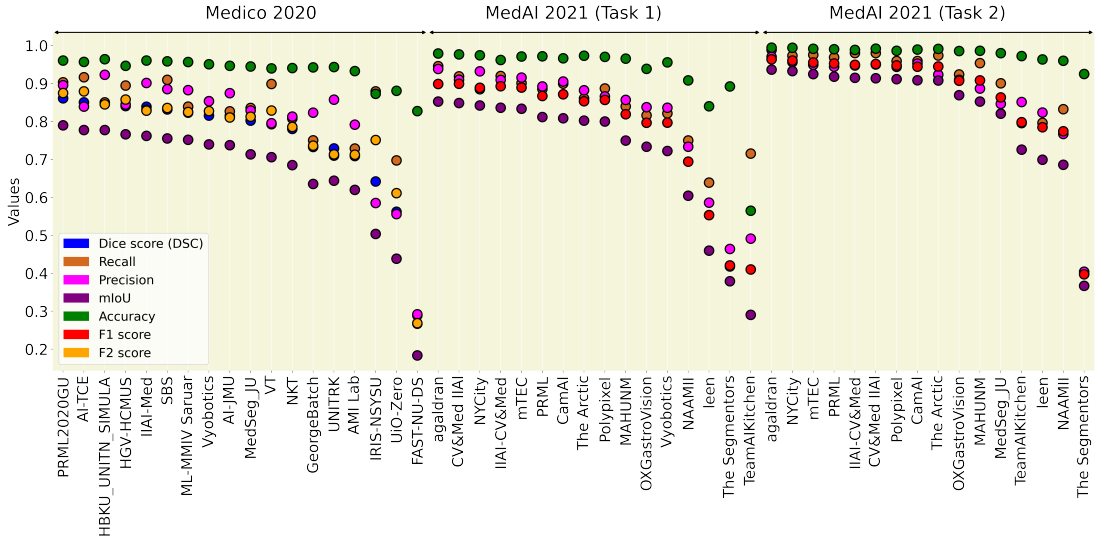
The rapid advancement in the AI-based techniques that support CAD systems has resulted in the introduction of numerous algorithms in the domain of medical image analysis, including colonoscopy. To assess the performance of these algorithms, it is important to benchmark on the particular set of datasets. It enables the comparison and analysis of different techniques and assists in identifying challenging cases that need to be targeted using improved methodologies. This also includes cases that are misled by the presence of artifacts and occlusion by surgical instruments [180]. Besides developing and analyzing



(a)



(b)



(c)

Figure 5.5: (a) Violin plots with overlaid swarm plots depicting statistics of submissions received for different tasks for the two challenges, (b) Dice score comparison of different teams in three tasks of Medico 2020 (polyp segmentation) and MedAI 2021 (Task 1: Polyp segmentation and Task 2: Instrument segmentation), and (c) Strip plots for all segmentation metrics (Dice score (DSC), recall, precision, mIoU score, accuracy, F1 score, and F2 score) reported by different teams in both challenges for all test data samples.

AI-based algorithms, it is crucial to include explainability and interpretability to infuse trust and reliance during the adoption of automated systems in clinical settings. Therefore, the challenges discussed in this chapter focus on lesion and instrument segmentation and emphasize the importance of transparency in medical image analysis. This section covers the findings, limitations, analysis of failing cases, trust, safety and interpretability of the methods, and future steps and strategies covering both challenges, Medico 2020 and MedAI 2021.

Medico 2020 Challenge Methods

Most of the methods reported in the Medico 2020 challenge focus on encoder-decoder architecture for example, U-Net, ResUNet++, PraNet, Efficient UNet, etc). Other networks used include conditional GAN and Faster R-CNN. The overview of the methods is provided in Table 5.1. For more detailed architectural information, we have also included the backbone and algorithm used by each team. Further, we also report the nature of the algorithm and the choice basis of evaluation, such as mIoU, DSC or FPS. Additionally, we provide information about the augmentation and hyperparameters, such as loss function and optimizers. It is noteworthy that all the top three teams "PRML2020GU", "HBKU_UNITN_SIMULA" and "AI-TCE" used the encoder-decoder architecture. Out of 17 participating teams, only three teams adopted some other architectures. Comparative analysis shows that the highest-scoring encoder-decoder network outperforms the GAN-based approach by a significant margin of 0.3517 in mIoU and 0.2989 in DSC score. Similarly, compared to the R-CNN-inspired networks (team "IRIS-NSYSU"), the best approach (team "PRML2020GU") achieves an improvement of 0.2863 in mIoU score and 0.2191 DSC score.

Medico 2020 challenges provide valuable insight and trends for the polyp segmentation and biomedical image analysis challenges. Most deep learning frameworks submitted for the challenge used the Adam optimizer to optimize their network. However, a handful of teams used other optimizers, such as SGD or RMSProp. Additionally, most of the teams used data augmentation to boost the number of training samples prior to training their frameworks to improve the performance of their architecture. There have been different preferences in loss function where most of the team used "BCE + DSC loss", "binary cross-entropy", IoU loss, etc. However, from the results of the top three teams, it can be concluded that "BCE + DSC loss" is best for this dataset. Similarly, in terms of the backbone for the model architecture, the EfficientNet variant (selected by PRML2020GU) or EfficientNetB4

(selected by AI-TCE) were most favorable.

MedAI 2021 Challenge Methods

The summary of the different approaches adopted by the participating teams of the MedAI2021 Challenge is presented in Table 5.2. To provide a brief overview of the general techniques adopted by the different teams, they can be categorized based on the nature of the approach followed, such as ensemble models, encoder-decoder based architectures, CNN, and hybrid CNN models. Almost all the teams presented the same model for both the tasks proposed in the challenge. Most teams explored ensemble modelling, encoder-decoder networks, or a combination of both in the polyp segmentation task. Another criterion of categorization could be CNN or transformed-based approaches. It is observed that the top-ranked team "agaldran" utilized two encoder-decoder networks and reported a mIoU score of 0.8522. Similarly, "CV&Med IIAI" was ranked second, and Team "NYCity" was ranked third in the polyp segmentation task with a competitive mIoU value of 0.8484 and 0.8418, respectively. Similar to the Medico 2020 polyp segmentation challenge, where GAN-based methods were adopted by teams (for example, Team "leen") failed to perform well in this challenge for polyp and instrument segmentation tasks. It is to be noted that the winning team, "agaldran" used a double encoder-decoder structure with two U-Net, where they incorporated FPN and Resnext101 as the pretrained decoder. They also used SAM and Adam optimizer to optimize the model further. The other competitive team "CV&Med IIAI" used the SINetv2 algorithm with PVTv2 as the backbone, and NYCity used the combination of HarDNet-85 ResNet101.

In the MedAI2021 instrument challenge, participants mainly focused on either ensemble models or encoder-decoder networks similar to the polyp segmentation task. As the majority of the teams utilized the same model that they proposed for the polyp segmentation problem in this task, the categorization of overall methods remains the same as that of the first task described above. The top rank is secured by Team "agaldran", with encoder-decoder architecture, pyramid network as the decoder, and Resnext101 as the pre-trained decoder. The second-ranked model by Team "NYCity" is the CNN and transformer based ensemble model, which achieved only a slight difference in the scores from the leading model. mTec was ranked third in the challenge, which used dual parallel reverse attention edge network (DPRA-EdgeNet) [168]. The architecture used HardNet [181] as the backbone.

The challenge shows that most of the teams were reluctant to share their method (refer

to Table 5.7). From the table, it can be seen that only five teams were qualified for the doctor evaluation. Additionally, the quality of the code submitted by most of the team was not satisfactory. Most of the participants did not put much effort into the readme file. Additionally, most teams neglected the failure analysis, ablation study and explainability in their submission. Moreover, based on the doctor's evaluation, only the solution provided by a few teams (for example, "agaldran", "mTEC" "CamAI", "The Arctic", and "IIAI-CV&Med") was considered useful and understandable.

Analysis of the Failed Cases

We have analyzed the regular and failing cases in polyp and surgical tool segmentation to highlight the limitations of the current methods so that these cases can be considered during further algorithm development. From the results on the test dataset, it was observed that most of the algorithms failed on diminutive and flat polyps located in the left colon. These are the challenging classes in the colon and require effective detection and diagnosis systems. Similarly, although most of the methods performed well on the diagnostic and therapeutic surgical tool, there were issues with the images having caps and forceps. This happens as algorithms could still struggle with difficult and rare cases like sessile polyps, even if they perform well on overall quantitative metrics. Therefore, investigating the cause for misclassification for such samples in the dataset and failure analysis will be critical to focus for future research. This can include generalization performance evaluation on unseen test data from different hospitals. Such investigations can reduce the chances of underperformance on rare cases.

Trust, Safety, and Interpretability of Methods

Integrating CAD in clinical settings necessitates addressing factors such as trust, safety, and interpretability to ensure its adoption. The high variations and potential bias in the curated datasets used to train such models and the actual scenarios in which they are adopted create a high chance of biases, impacting the generalizability of the method. Such bias ultimately makes it challenging to infuse trust while adopting CAD tools and questions the safety of patients. To tackle this issue, we introduced a transparency task in the MedAI2021 challenge that underscores the need for interpretability, reproducibility, and explainability in medical AI research, including polyp and instrument segmentation.

Our initiative aimed to light the potential risk that can arise from wrong decisions based on model and algorithmic bias. Our dataset contained polyp cases with varied ap-

pearances in terms of shapes, sizes, the presence of artifacts, lightning conditions, textures, and the different numbers of polyps per image that are encountered in real-world clinical settings. Additionally, we have included frames containing surgical instruments to support the cases of occluded endoluminal elements or polyps that could arise in general. Some of the methods adopted by the participating teams include the submission of intermediate heatmaps using approaches like layer-wise relevance propagation that showed visual explanation and highlighted the model decision-making process. Team "agaldran" provided detailed ablation studies in support of the predictions obtained. By promoting transparency through subjective analysis and addressing potential biases, the MedAI challenge aimed to foster trust in the presented solution and ensure safety in adopting such methods in the clinic.

Limitation of the Medico 2020 and MedAI 2021

In our study, we aimed to standardize the challenge of polyp and instrument segmentation by providing the same test sets and evaluation metrics to all participants. To achieve this, we introduced variable polyp cases, including polyps with different sizes, noisy frames with artifacts, blurry images, and occlusion. We also added regular frames to the test set to ensure that participants drew the ground truth manually and did not cheat. However, our study has some limitations. Although we used datasets collected from four medical centers in Norway, these images are from a single country, limiting the ethnicity variance though there are very limited differences, if any, in the mucosal appearance between ethnicities. Nevertheless, there is a need for a more diverse dataset that includes multiple ethnicities and countries also because the prevalence of various diseases varies between regions. Moreover, the current models should be tested on multi-center datasets to assess their generalization ability.

There was no online leaderboard in our challenge due to the Mediaeval policy. Therefore, we manually calculated the predictions submitted by each team. Each team had limitations of 5 submissions for each task, which restricted further optimization opportunities. Although we have also introduced normal findings from the GI tract to trick the participants and models, our challenge only used still frames and did not incorporate video sequence datasets. Even when the best-performing algorithms are tested on a temporal video sequence dataset, it is possible that the performance can drop. Most of the images are only from WLI. Although our dataset was annotated by one annotator and checked by two gastroenterologists, there is still a possibility of bias in the labels. In the accessory in-

strument challenge, we had more images from the stomach class than accessory instruments such as biopsy forceps or snares due to the lack of availability of datasets. Finally, despite including diverse cases in the polyp and instrument segmentation challenge, we still had limited flat and sessile polyps, frequently missed during routine colonoscopy examinations. Incorporating multi-center data and video sequences data and addressing label biases will lead to more comprehensive and reliable evaluations of AI-based colonoscopy systems.

Future Steps and Strategies

In our study, we aimed to promote transparency and interpretability in machine learning models for the GI tract setting. However, more work is needed to understand how decisions are made and identify potential biases or errors in a quantitative manner to build trust in such systems in a clinical setting. To achieve this, we plan to test the best-performing algorithms on large-scale datasets to observe their scalability. We will consider metrics that weigh speed, accuracy, and robustness for better objective assessments and introduce more distance-based metrics such as Hausdorff distance and normalized surface distance for improved fairness.

We will emphasize more transparent decision-making methods and visualize interpretability results while focusing on clinical relevance rated by expert clinicians instead of just one objective metric. To achieve this, we have already started collecting large-scale datasets and plan to build a tool if the algorithms are robust enough and verified by our gastroenterologists. Next, we will propose a challenge to polyp video sequences analysis. We will explore the integration of state space models, such as Video Vision Mamba-based framework [182], to capture the temporal information in video sequences that affect the efficiency and accuracy of segmentation tasks. It is worth noting that there has been innovation within hardware (colonoscope) for safer medical colonoscopy devices, such as developing fully flexible automated colonoscopes to offer expanded fields of view rather than $120 - 170^{\circ}$ visualization, which can capture dead spots, improving the lesions' miss-rate. These scopes are currently in the final stage of development. This hardware would require high processing speed to locate potential lesions in real time for a smooth workflow. We believe these solutions from our challenge could help address the complexities with the improved hardware and improved image quality.

5.3 GastroVision

Here, we provide detailed information about the dataset, acquisition protocol, ethical and privacy aspects of data and suggested metrics.

5.3.1 Dataset Details

GastroVision is an open-access dataset that incorporates 8,000 images pertaining to 27 different labeled classes (Fig. 5.6). Most image are obtained through WLI, while a few samples are acquired using NBI. These classes are categorized into two broad categories: *Upper GI tract* and *Lower GI tract*. The number of images under each class is presented in Fig. 5.7. These classes indicate findings acquired from the GI tract. It can be observed that the sample count is not balanced across classes, which is generally experienced in the medical image acquisition process as some findings occur less often. Releasing these classes in the dataset will allow the researchers to leverage the fast-emerging AI techniques to develop methods for detecting such rare but clinically significant anomalies. All the images are stored in JPG format, and the overall size is around 1.8 GB. The resolution details of the images can be found in Fig. 5.8. *GastroVision* is provided as a structured directory, with each class having a specific folder. For example, the '*Accessory tools*' folder contains all images featuring diagnostic and therapeutic tools.

Upper GI Tract:

Upper GI endoscopy examines the esophagus, stomach, and duodenum. The various classes covered in this GI tract are discussed below as three subcategories: *normal findings*, *anatomical landmarks*, and *pathological findings*. A detailed categorization is shown in Fig. 5.6. The **normal stomach** serves as a critical site for initial digestion, while the **duodenal bulb**, the first part of the small intestine, is critical for nutrient absorption. Anatomical landmarks are used as reference points to indicate a specific location and assist in navigating during endoscopy procedures. The **gastroesophageal junction** is an anatomical area where **esophagus** joins the **stomach** also alining to the **normal z-line**, a transitional point where the esophagus's squamous epithelium and the stomach's columnar mucosa lining join. **Pylorus** is a sphincter connecting the stomach and the duodenum, the first part of the small intestine.

Apart from these anatomical landmarks, any pathological conditions may be encountered during endoscopy. **Esophagitis**, the most common abnormality, is characterized by

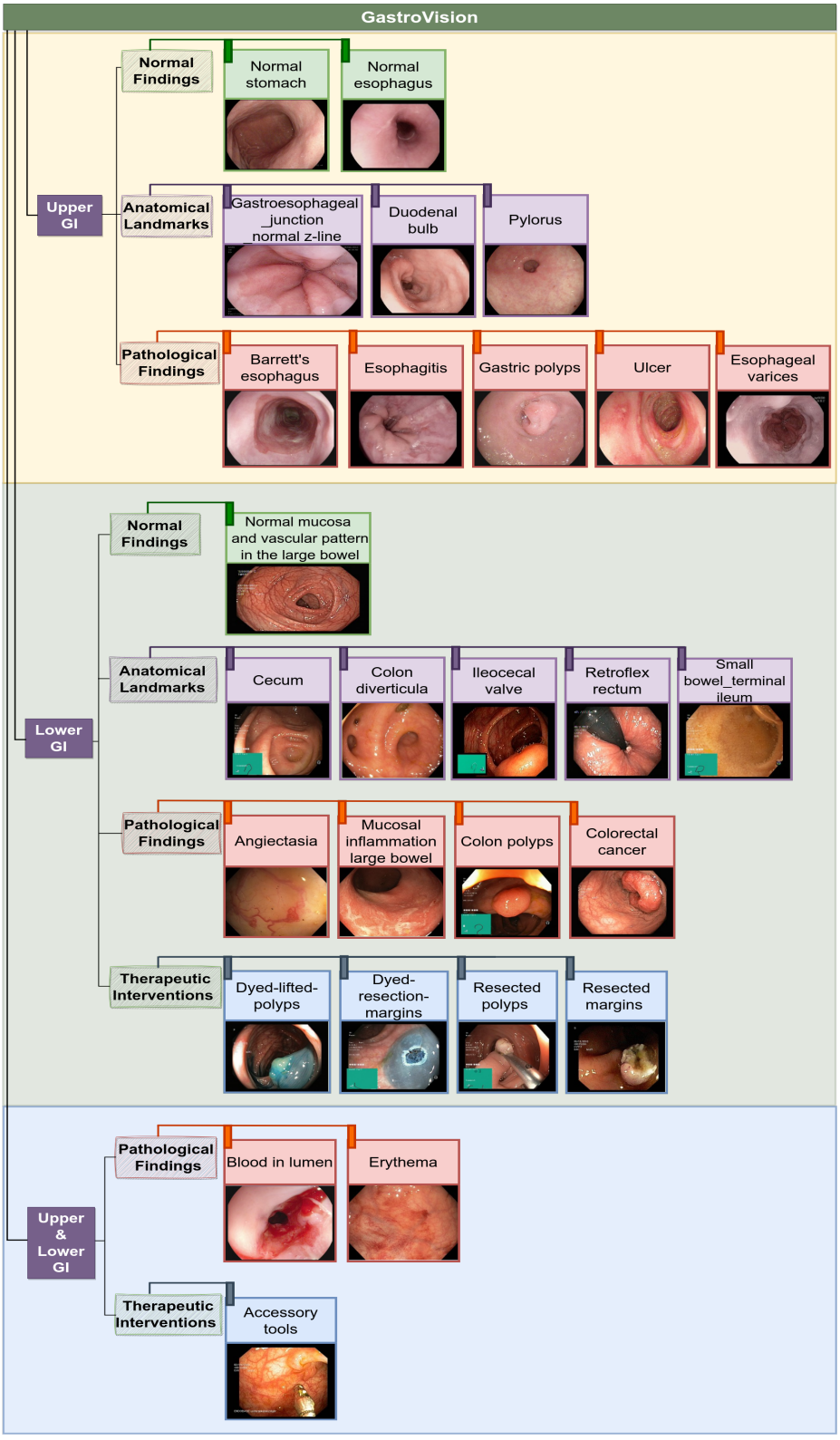


Figure 5.6: Example images from the gastrointestinal tract showing distinct findings from the upper and lower GI tract.

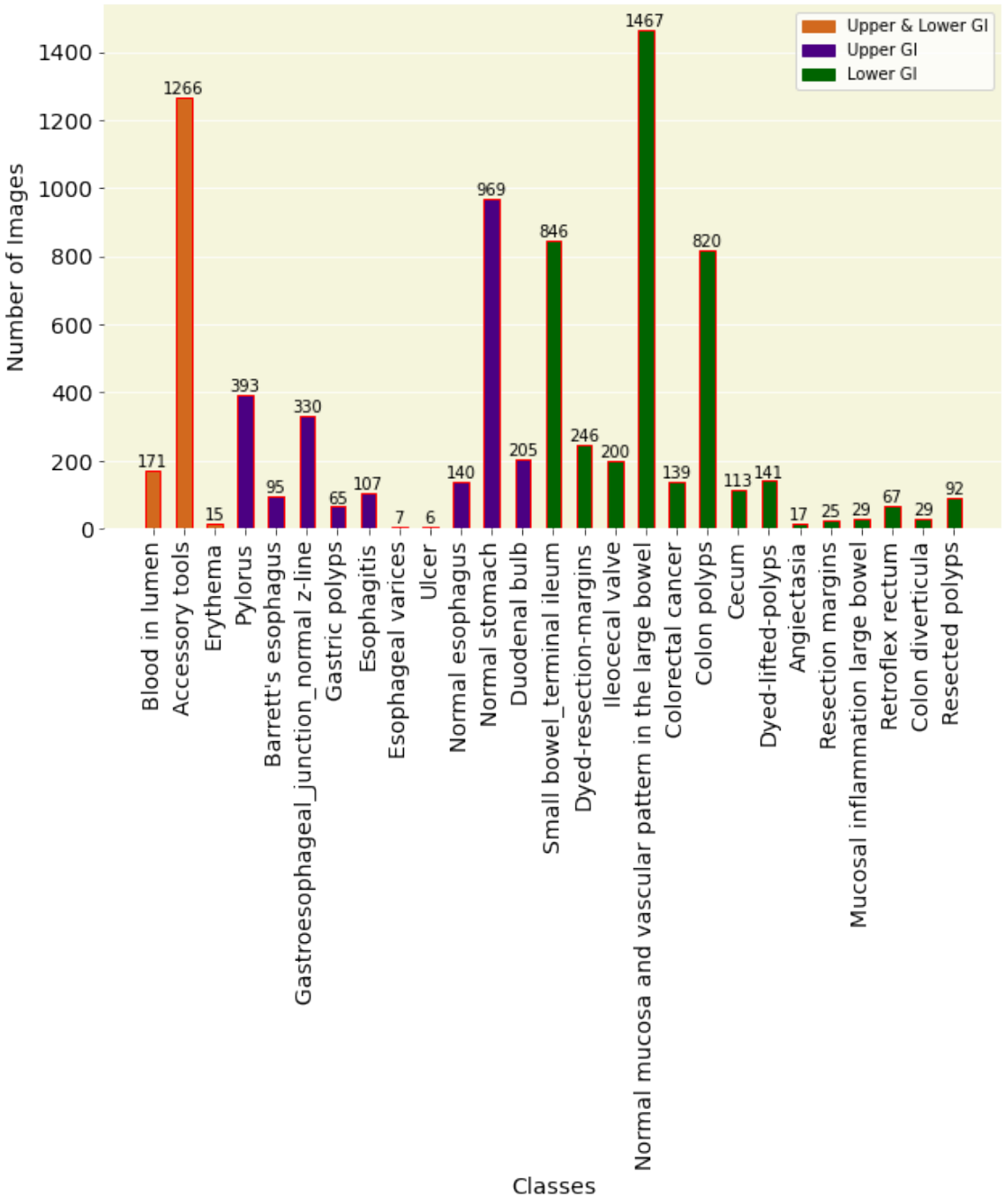


Figure 5.7: The figure shows the number of images per class. Some classes have few samples because of the rarity of the findings and the technical challenges associated with obtaining such samples in endoscopic settings.

an inflammation of the esophagus. This disease is graded based on its severity according to the Los Angeles classification. For example, grade B refers to the condition when the mucosal break is limited to the mucosal fold and is more than 5 mm long. In grade D,

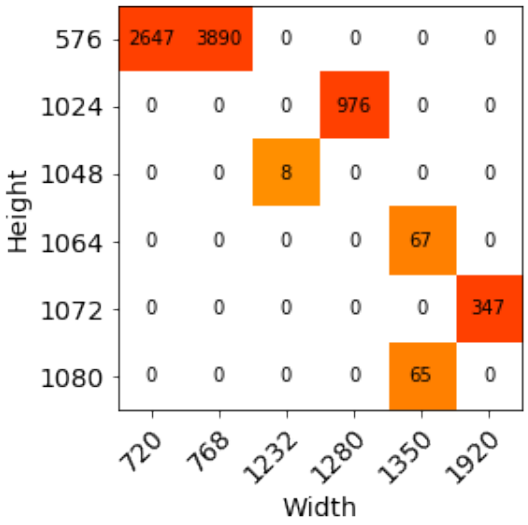


Figure 5.8: Resolutions of the 8,000 images of GastroVision.

mucosal break affects 75% of the esophageal circumference. Long standing esophagitis may cause **Barett's esophagus**, a condition in which the cells of the esophagus's lining start to change, and tissues appear red. This is a precancerous condition. Other frequent lesions observed are **polyps**, abnormal tissue growth or ulcers. **Gastric polyps** are abnormal growths in the stomach lining. **Ulcers** are the open sores in the stomach or duodenum that can lead to discomfort and bleeding. **Esophageal varices** result from portal hypertension, causing swollen veins in the esophagus. **Erythema** refers to redness, often indicating inflammation and **blood in the lumen** denotes bleeding. **Accessory tools** aid in investigating and diagnosing upper and lower GI tract conditions for targeted treatment.

Lower GI Tract:

The lower GI tract is examined by colonoscopy to investigate any abnormalities in the colon, the rectum, and the terminal ileum (the last part of the small bowel). Here, we covered one more subcategory, *therapeutic interventions*, in addition to *normal findings*, *anatomical landmarks*, and *pathological findings*. A detailed class-wise division is shown in Fig. 5.6.

The **normal mucosa and vascular pattern in the large bowel** is essential for absorbing water and electrolytes. The different anatomical landmarks associated with lower GI include **cecum** (first part of the large intestine), visualizing the appendiceal orifice, **ileocecal valve** (sphincter muscle between ileum and colon), and the **small bowel**. During the colonoscopy, these anatomical landmarks act as reference points to prove complete examination. Retroflexion in the rectum is performed to visualize a blind zone, using

the bending section of the colonoscope to visualize the distal area of the colon, called **rectroflex-rectum**. The **terminal ileum**, the last part of the small intestine, aids in nutrient absorption. **Colon diverticula**, small pouch-like protrusions, can form along the colon's weakened wall, often in the sigmoid colon [183].

During the colonoscopy, the endoscopist navigates through these landmarks and looks for abnormalities such as **polyps**, **angiectasia**, and inflammation like **ulcerative colitis**. **Angiectasia** is a common lesion representing abnormal blood vessels and is responsible for obscure recurrent lower GI bleeding. These can easily be distinguished from the **normal vessels** shown in Fig. 5.6. **Colorectal cancer** occurs in the colon or rectum. One of the early signs of this colorectal cancer can be detected through **colon polyps**. **Mucosal inflammation in the large bowel** may be caused by different factors, such as infections or chronic inflammatory conditions.

Apart from the aforementioned pathological conditions, several therapeutic interventions are adopted to treat the detected anomalies effectively. It frequently involves the removal of the lesion/polyp. The surrounding of such **resected polyps**, also called the **resection margins** or resection sites, are then considered for biopsies. To enhance lesion demarcation, a solution containing indigo carmine is injected, making resection easier. The appearance of blue color underneath the **dyed-lifted-polyp** provides accurate polyp margins. After resecting such polyps, the underlying region, known as **dyed-resection-margin**, appears blue. These margins are important to examine for any remaining tissue of the resected polyp.

5.3.2 Dataset Acquisition, Collection and Construction

Data Acquisition and Collection:

The dataset images are acquired from two centers (Department of Gastroenterology, Bærum Hospital, Vestre Viken Hospital Trust (VV), Norway and Karolinska University Hospital, Stockholm, Sweden) using standard endoscopy equipment from Olympus (Olympus Europe, Germany) and Pentax (Pentax Medical Europe, Germany). A team of expert gastroenterologists, one junior doctor, and two computational scientists were involved in the labelling of the images and the related review process. It is worth noting that for dataset collection, we labeled some of the unlabeled images from the HyperKvasir dataset and included them in our dataset. Additionally, we labeled the images acquired from the Karolinska University Hospital to their respective classes for developing a diverse and multi-center "GastroVision" dataset.

Ethical and Privacy Aspects of the Data:

The dataset is constructed while preserving the patients' anonymity and privacy. All videos and images from Bærum hospitals were fully anonymized, following the GDPR requirements for full anonymization. Hence, it is exempted from patient consent. The files were renamed using randomly generated filenames. The Norwegian Privacy Data Protection Authority approved this export of anonymized images for research purposes. As the dataset development procedure involved no interference with the medical treatment or care of the patient, it has also been granted an exemption for approval by the Regional Committee for Medical and Health Research Ethics - South East Norway. Similarly, the data collection process at Karolinska University Hospital, Sweden, is completely anonymized as per the GDPR requirements.

5.3.3 Suggested Metrics

Standard multi-class classification metrics, such as Matthews Correlation Coefficient (MCC), micro and macro averages of recall/sensitivity, precision, and F1-score, can be used to validate the performance using our dataset. MCC provides a balanced measure even in cases with largely varying class sizes. A macro-average will compute the metric independently for each class and then take the average, whereas a micro-average will aggregate the contributions of all classes to compute the metric. Recall presents the ratio of correctly predicted positive observations to all the original observations in the actual class. Precision is the ratio of correctly predicted positive observations to all the positive predicted observations. F1-score integrates both recall and precision and calculates a weighted average/harmonic mean of these two metrics.

5.3.4 Experiments and Results

In this section, we describe the implementation details, technical validation and the limitation of the dataset.

Implementation Details

All deep learning diagnostic models are trained on NVIDIA TITAN Xp GPU using PyTorch 1.12.1 framework. A stratified sampling is performed to preserve the similar distribution of each class during 60:20:20 training, validation, and testing split formation. The images are resized to 224×224 pixels, and simple data augmentations, including random rotation and

Table 5.8: Results for all classification experiments on the Gastrovision dataset.

	Macro Average			Micro Average			
Method	**Prec.**	**Recall**	**F1**	**Prec.**	**Recall**	**F1**	**MCC**
ResNet-50 [118]	0.4373	0.4379	0.4330	0.6816	0.6816	0.6816	0.6416
Pre-trained ResNet-152 [118]	0.5258	0.4287	0.4496	0.6879	0.6879	0.6879	0.6478
Pre-trained EfficientNet-B0 [184]	0.5285	0.4326	0.4519	0.6759	0.6759	0.6759	0.6351
Pre-trained DenseNet-169 [113]	0.6075	0.4603	0.4883	0.7055	0.7055	0.7055	0.6685
Pre-trained ResNet-50 [118]	0.6398	0.6073	0.6176	0.8146	0.8146	0.8146	0.7921
Pre-trained DenseNet-121 [113]	**0.7388**	**0.6231**	**0.6504**	**0.8203**	**0.8203**	**0.8203**	**0.7987**

random horizontal flip, are applied. All models are configured with similar hyperparameters, and a learning rate of $1e^{-4}$ is initially set with 150 epochs. An Adam optimizer is used with the *ReduceLROnPlateau* scheduler. More description about the implementation details and dataset can be found on our GitHub page [6].

Technical Validation

To evaluate the presented data for technical quality and classification tasks, we performed a series of experiments using some state-of-the-art deep learning models. The purpose of this preliminary validation is to provide baseline results that can be referred to for comparison by future researchers. We carried out multi-class classification using CNN-based models, namely, ResNet-50 [118], ResNet-152 [118], EfficientNet-B0 [184], DenseNet-121 [113], and DenseNet-169 [113], considering their competent performance in GI-related image-based tasks in the literature [185]. Note that we have only included classes with more than 25 samples in the experiments, which resulted in 22 classes in total. However, we also release the other classes with fewer samples to welcome new interesting findings in areas similar to one-shot learning.

The different experiments performed include *(a) ResNet-50*: The model is randomly initialized, and an end-to-end training is done, *(b) Pre-trained ResNet-50* and *(c) Pre-trained DenseNet-121*: The models are initialized with pre-trained weights, and then all layers are fine-tuned, *(d) Pre-trained ResNet-152*, *(e) Pre-trained EfficientNet-B0* and *(f) Pre-trained DenseNet-169*: The models are initialized with pre-trained weights, and only the updated last layer is fine-tuned. All the above pre-trained models use ImageNet weights. The associated results are shown in Table 5.8. It can be observed that the best outcome is obtained using the pre-trained DenseNet-121. A class-wise analysis using the same model is provided in Table 5.9 and Fig. 5.9. It shows that while most classes achieved satisfactory prediction outcomes, a few proved to be very challenging for the classification model. For

[6] https://github.com/DebeshJha/GastroVision

Class	Precision	Recall	F1-score	Support
Accessory tools	0.93	0.96	0.95	253
Barrett's esophagus	0.55	0.32	0.4	19
Blood in lumen	0.86	0.91	0.89	34
Cecum	0.33	0.17	0.23	23
Colon diverticula	1	0.33	0.5	6
Colon polyps	0.78	0.87	0.82	163
Colorectal cancer	0.63	0.41	0.5	29
Duodenal bulb	0.72	0.76	0.74	41
Dyed-lifted-polyps	0.86	0.86	0.86	28
Dyed-resection-margins	0.94	0.92	0.93	49
Esophagitis	0.5	0.23	0.31	22
Gastric polyps	0.6	0.23	0.33	13
Gastroesophageal_junction_normal z-line	0.65	0.85	0.74	66
Ileocecal valve	0.74	0.7	0.72	40
Mucosal inflammation large bowel	1	0.33	0.5	6
Normal esophagus	0.72	0.82	0.77	28
Normal mucosa and vasular pattern in the large bowel	0.81	0.87	0.84	293
Normal stomach	0.9	0.86	0.88	194
Pylorus	0.8	0.92	0.86	78
Resected polyps	0.33	0.11	0.17	18
Retroflex rectum	0.75	0.43	0.55	14
Small bowel_terminal ileum	0.86	0.85	0.85	169

Table 5.9: Class-wise performance associated with the best outcome obtained using pre-trained DenseNet-121.

a more detailed analysis, we plotted a two-dimensional t-SNE embedding for *GastroVision* (Fig. 5.10). The classes like *Normal stomach*, *Dyed-resection-margins*, which present a clear distinction in the t-SNE embedding, are less often misclassified. The above points could be the reasons for the F1-score of 0.88 and 0.93 in the case of *Dyed-resection-margins* and *Normal stomach* classes, respectively. On the other hand, there are some overlapping classes such as *Cecum* and *Normal mucosa and vascular pattern in the large bowel* or *Colorectal cancer* and *Colon polyps* which do not present clear demarcation with each other and hence, are likely to be misclassified.

Considering the overall results and many overlapping classes (without distinct clustering), it can be inferred that classifying GI-related anatomical landmarks and pathological findings is very challenging. Many abnormalities are hard to differentiate, and the rarely occurring findings have higher chances of getting misclassified. This presents the challenge of developing a robust AI system that could address multiple aspects important for GI image classification, e.g., many findings are subtle and difficult to identify, and some findings

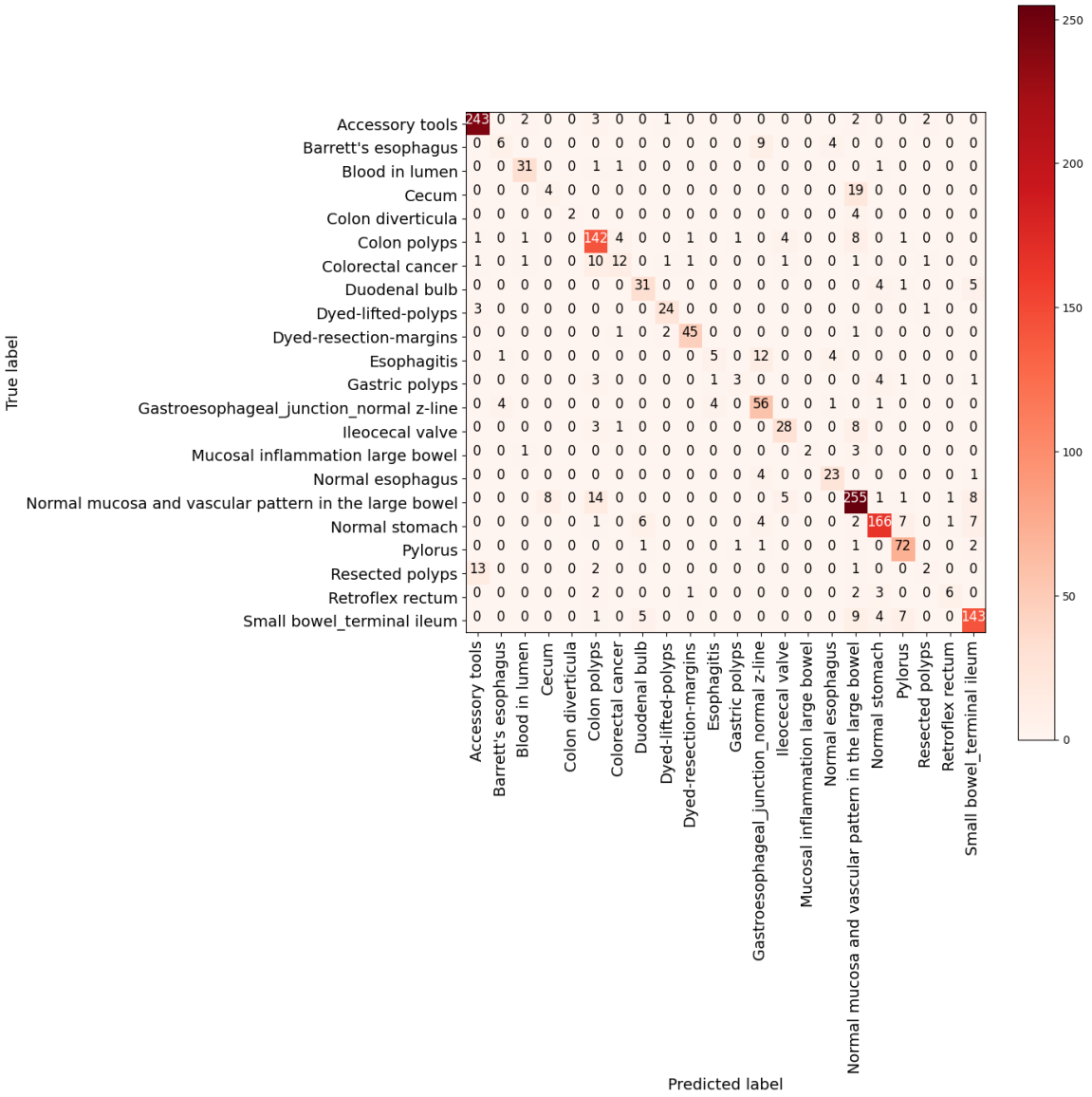


Figure 5.9: Confusion matrix for the best outcome obtained using pre-trained DenseNet-121.

are not easily acquired during the endoscopy procedure, which results in less number of data samples. Such underrepresented classes need to be explored with some specific algorithms specially designed to leverage the availability of a few hard-to-find samples. Thus, the potential of the baseline results and associated issues and challenges motivate the need to publish this dataset for further investigations.

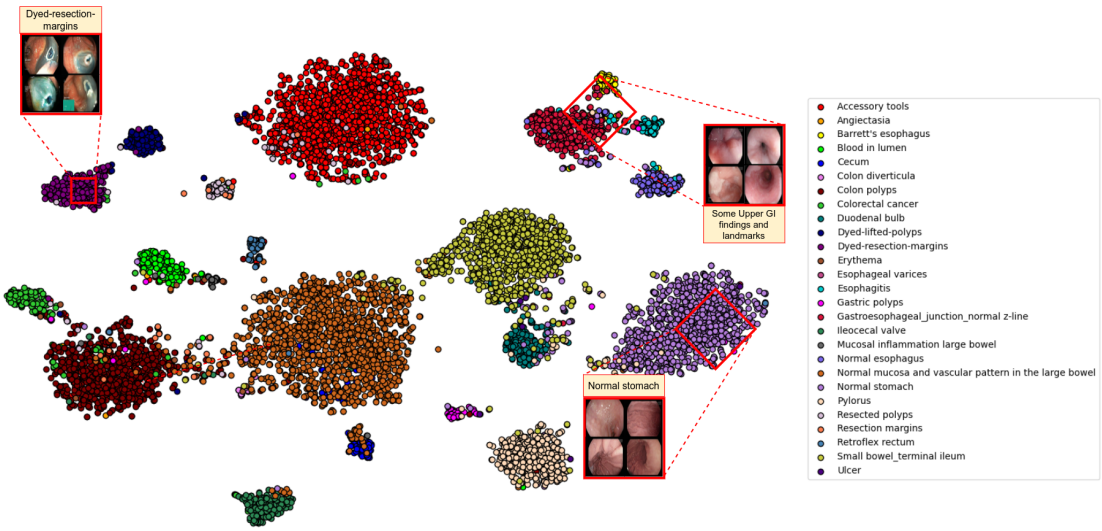


Figure 5.10: Two-dimensional t-SNE embedding for GastroVision. The pre-trained DenseNet-121 model, which is further trained on our training set, is used to extract features. Some sample images are shown with either a specific or a broader (due to multiple overlapping classes) categorization.

Limitation of the Dataset

Our dataset, *GastroVision*, is a unique and diverse dataset with the potential to explore a wide range of anatomical and pathological findings using automated diagnosis. Although this labeled image data can enable the researchers to develop methods to detect GI-related abnormalities and other landmarks, the current version lacks segmented annotations, which could further enhance the treatment experience and surgical procedures. It is important to note that some classes (for example, colon diverticula, erythema, cecum, esophagitis, esophageal varices, ulcer and pylorus) have only a few images. Despite this limitation, our dataset is well suited for one-shot and few-shot learning approaches to explore some GI-related conditions that have still not received attention in medical image analysis. In the future, we plan to extend the dataset by including more classes and a larger number of samples, along with ground truth for some of the classes that could be used for segmentation purposes as well as images with higher resolution from the most recent endoscopy systems.

5.4 Chapter Summary

In this chapter, we presented a case study with the aim of providing a comprehensive analysis of the methods used by participants in the Medico 2020 and MedAI 2021 competitions for different medical image analysis tasks. We designed the tasks and datasets to demonstrate that the best-performing approaches were relatively robust and efficient for automatic polyp and instrument segmentation. We evaluated the challenge based on several standard metrics. In MedAI 2021, we also used a quantitative approach, where a multi-disciplinary team, including gastroenterologists, accessed each submission and evaluated the usefulness and understandability of their results. Through the qualitative results, we found that more generalizable and transparent methods are needed to be integrated into real-world clinical settings. During the "performance task" and "algorithm efficiency" tasks, we observed a trade-off between accuracy and inference time when tested across unseen still frames. For the instrument segmentation challenge, we observed that almost all teams performed relatively well, as segmenting instruments is easier than polyp segmentation. From the transparency task, we observed that more effort is required from the community to enhance the transparency of their work. Overall, we also observed that several teams demonstrated the use of data augmentation and optimization techniques to improve performance on specific tasks. Our study highlights the need for multi-center dataset collection from larger and more diverse populations, including experts from various clinics worldwide.

Further, we presented a new multi-class endoscopy dataset, *GastroVision*, for GI anomalies and disease detection. We have made the dataset available for the research community along with the implementation details of our method. The labeled image data can allow researchers to formulate methodologies for classifying different GI findings, such as important pathological lesions, endoscopic polyp removal cases, and anatomical landmarks found in the GI tract. We evaluated the dataset using some baseline models and standard multi-class classification metrics. The results motivate the need to investigate specific techniques for GI-related data better. Having a diverse set of categories labeled by expert endoscopists from two different centers, *GastroVision* is unique and valuable for computer-aided GI anomaly and disease detection, patient examinations, and medical training.

6

Synthetic Dataset Creation: Towards Improved Polyp Segmentation and Classification

Chapter Highlights

- We address the issues associated with medical data acquisition, including legal restrictions and manual effort for acquisition and annotations.
- We develop two methods to generate synthetic medical data that eliminate the need to undergo long data acquisition and annotation procedures.
- The generated data can be controlled using two mechanisms: image-based and text-based input controls.
- We also introduce the concept of cross-class labels that allow learning features from other classes and generate images from combinations of text prompts without additional labels.
- This chapter is based on the publications "*ControlPolypNet: Towards Controlled Colon Polyp Synthesis for Improved Polyp Segmentation*", presented at the CVPR (DCAMI) 2024 and "*Generating Diverse Modality Colonoscopy Images Leveraging Cross-Class Labels*", currently under review.

This chapter addresses the issue of expensive manual effort and privacy concerns associated with medical data acquisition and annotation. Such hurdles result in medical data scarcity, impacting the performance of deep learning methods as it depends on the quality and number of samples presented during training. Existing lesion detection and diagnosis tasks of a CAD system, such as polyp segmentation and classification methods, typically rely on conventional augmentation techniques such as rotation, flipping, etc. Simply relying on traditional augmentation techniques restricts the scale-up of the dataset to a certain extent, depending on the dataset size, and limits diversity among samples. Moreover, the existing classification techniques fail to capture different imaging modalities (NBI/WLI) that are paired with colonoscopy procedures for accurate classification of two major types of polyps: adenomatous (malignant potential) and hyperplastic (benign). The reason for such failures stems from the same issue of diverse data unavailability in the medical domain. To overcome these issues, synthetically generated images can be utilized to complement the hard-to-obtain annotated authentic medical data. Recently, generative artificial intelligence has been gaining prominence in this domain. Additionally, various generation-controlling mechanisms using text prompts and images have been introduced to obtain visually appealing and desired outcomes in a better-controlled manner. Therefore, we develop two frameworks; one explores the image-based control mechanism, and the other framework explores the text-controlled generation mechanism. Our first model, ControlPolypNet is a novel, stable diffusion based framework. We control the generation process (polyp size, shape and location) using a novel custom-masked input control, which generates images preserving important endoluminal information. Additionally, our model comprises a detection module, which discards some of the generated images that do not possess lesion-characterizing features, ensuring clinically relevant data. We further utilize the generated polyp frames to improve performance in the downstream task of polyp segmentation. Using these generated images, we found an average improvement of ***6.84%*** *and* ***1.3%*** *(JI) on CVC-ClinicDB and Kvasir-SEG, respectively, in the polyp segmentation task. Further, we develop our second model, PathoPolyp-Diff, that generates text-controlled synthetic im-*

*ages with diverse characteristics in terms of pathology, imaging modalities, and quality. In the process, we leverage the cross-class labels to make the model learn features from other classes and, hence, reduce the burdensome task of data annotation. We validate the effectiveness of text-controlled synthesis and cross-class label learning by performing polyp classification (adenomatous/hyperplastic) with different imaging modalities (NBI/WLI) and text prompts. The experimental results report an improvement of up to **7.91%** in BA using a publicly available dataset. Moreover, the cross-class label learning achieves a statistically significant improvement of up to **18.33%** in BA during video-level analysis.*

6.1 Introduction

In this chapter, we propose two frameworks that focus on generating synthetic medical data, thus overcoming the long procedures involved in clinical data acquisition and labeling. The dataset curated and designed in our previous chapter involved expert gastroenterologists and computational scientists for the labelling and reviewing procedures. These procedures took a long time and extensive effort. Moreover, one must be cautious about annotation, as labelling medical data is critical. Considering these scenarios, one possible solution is to expand the training data by incorporating synthetic data. This solution is viable and offers various benefits: *(1) It does not require the time-consuming task of manual labelling. (2) It eliminates the long process of obtaining data privacy informed consent, accelerating dataset development. (3) It provides an opportunity to obtain hard-to-find anomalies that are difficult to observe during routine colonoscopy.*

To generate realistic-looking synthetic data, in recent years, GANs have been widely used in various fields, including medical imaging [102, 111]. Despite the improved performance in the downstream tasks, the issue of convergence instability of GANs and their limited contributions in such tasks resulted in the development of currently trending diffusion models [186, 187]. Diffusion models are expected to generate more realistic images and support text-to-image generation, thus facilitating automated systems with text prompts for better control. These models have been explored in many medical applications, such as image-to-image translation [188], reconstruction [189], image generation [190], segmentation [191], and classification [192], especially using radiology images. However, colonoscopy images are not much explored and require validations on the diffusion models' ability to

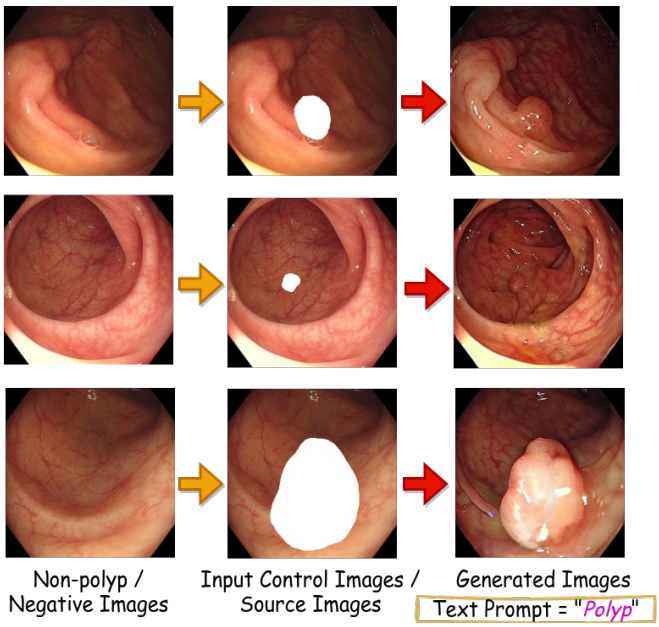


Figure 6.1: ***Controlling*** polyp generation using ***custom masks*** while leveraging largely accessible non-polyp/negative images. We turned negative samples into positive ones with controlled polyp shape, size and location, simultaneously enforcing diversity in generated samples.

learn and generate complex patterns. Besides visually satisfactory image formations, these models must be evaluated on their ability to retain clinically significant information and the usefulness of generated data for downstream tasks such as polyp segmentation.

In this chapter, we develop two diffusion based frameworks that generate synthetic colonoscopy images based on two different control mechanisms: a) image-based and b) text-based. Considering the former control mechanism, we propose *ControlPolypNet*, based on ControlNet [193] architecture to generate realistic-looking polyp frames. The framework has a novel input control map, which converts non-polyp frames with normal mucosa to polyp frames. This process is summarized in Fig. 6.1. Additionally, we employ a detector in *ControlPolypNet* that discards frames that do not carry lesion-characterizing features. We evaluate the generated frame's quality in preserving the endoluminal scenes by calculating Frechet Inception Distance (FID), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index Measure (SSIM) metrics. Also, we evaluated the impact of the generated data on downstream tasks of polyp segmentation. Our method offers a more practical approach to data augmentation, which is expected to represent clinically relevant data with diverse characteristics.

Using the second control mechanism, we propose PathoPolyp-diff, a novel diffusion

based framework that generates a wide range of colonoscopy image types covering different pathologies, imaging modalities and quality. This model aims to assist endoscopists in better-automated diagnosis, as the current methods mainly rely on a single imaging modality. A generative model, which can generate a wide range of image classes, generally requires a well-designed dataset with comprehensive annotations that are essential for training. This ensures they can grasp intricate patterns linked to polyp-characterizing features and their associated pathological conditions. However, obtaining labels for each subtask, considering pathology, quality, and imaging modality, could be significantly expensive. Therefore, in this framework, we develop a method to perform cross-class label learning that helps leverage annotations from other classes and produce synthetic images representing a combination of text prompts. This approach ensures obtaining synthetic polyp images with different pathological characteristics (adenoma/hyperplastic) captured with different imaging techniques (WLI/NBI). This diverse set is obtained while maintaining the quality, overcoming the general artifacts present in a colonoscopy video: ghost colors, motion blur, low illumination, and fecal depositions.

6.1.1 Contributions

The main contributions of this chapter are summarized below:

- **ControlPolypNet (A diffusion model with image-based control mechanism)**
  - **Framework with novel user-configurable input control map:** We propose a novel approach using user-configurable input control to generate polyps while leveraging the largely accessible non-polyp frames. This control map can control the endoluminal objects and polyp generation (in terms of shape, size and location) using customized masks and non-polyp frames.
  - **Additional examination to avoid irrelevant synthetic information:** We employ a detector that verifies the quality of generated polyps and selects clinically appropriate synthetic polyps that carry lesion-characterizing features. The detector eliminates the risk of adding noise and irrelevant information to the generated data.
  - **Improved polyp segmentation performance:** We report enhanced polyp segmentation performance by augmenting two publicly available datasets using our synthetic images. This has been achieved without additional expensive manual annotation requirements.

- **PathoPolyp-Diff (A diffusion model with text-based control mechanism)**
  - **A novel approach to generate a diverse set of colonoscopy images:** We develop *PathoPolyp-Diff*, a novel model to generate text-controlled synthetic images that cover a wide range of categories, including pathology, imaging modalities, and quality. It can generate adenomatous and hyperplastic polyps combined with the desired imaging modalities, including NBI and WLI.
  - **Introduced cross-class label learning:** We introduce the concept of cross-class label learning that allows the model to learn labels from different classes, hence expanding the diversity of data generation and reducing the cumbersome task of dataset annotation.

6.1.2 Preliminaries

The architecture of the two proposed approaches is inspired by the concepts of Stable Diffusion Models (SD) and ControlNet. These models are discussed in detail below.

Stable Diffusion Models (SD): SD is a text-to-image model built upon the basic functionality of Latent Diffusion Models (LDM) [194]. It has been introduced to circumvent the issue of the high computational requirements of standard Diffusion Models (DM) [186]. This improvement is achieved by executing the diffusion process on latent space instead of pixel level using an autoencoding procedure. To understand the advantages of using LDM over DM, some basic details about the two concepts are given below.

Standard DMs: The standard DMs follow a parameterized backward process of a fixed Markov Chain to gradually denoise a noisy image a_t. It acts as a sequence of autoencoders $\epsilon_\theta(a_t, t)$ that serves as a denoising framework to predict the denoised version of a_t. Here, t is uniformly sampled between $[1, T']$, and T' denotes the noise steps. The related objective can be defined in a refined form as:

$$L_{DM} := \mathbb{E}_{a,\epsilon,t}[\|\epsilon - \epsilon_\theta(a_t, t)\|_2^2] \tag{6.1}$$

where $\epsilon \sim \mathcal{N}(0, 1)$.

Standard LDMs: These models leverage the ability of encoder and decoder architectures to represent significant information in compressed form and reconstruct it back in its original form. Such an attempt to use latent space enables the model to focus on important semantic details and perform efficiently with low computational resources. LDMs also keep track of time steps t_i, further embedded with the U-Net architecture. However, instead of

directly using an RGB image $a \in \mathbb{R}^{H \times W \times 3}$, it is processed through encoder E to obtain a_l. This representation is reconstructed into a'_l using a decoder D. The corresponding objective is given below:

$$L_{LDM} := \mathbb{E}_{E(a),\epsilon,t}[\|\epsilon - \epsilon_\theta(a_{lt}, t)\|_2^2] \tag{6.2}$$

The SD further utilizes a text encoder, which is a pre-trained CLIP [195]. It allows encoding the text prompts into embeddings. These text embeddings are then fused with the encoder and decoder of U-Net using cross-attention layers. This cross-attention mechanism helps condition the model using a text prompt b after processing the embeddings through an encoder $\mathcal{Z}$.

$$L_{LDM_b} := \mathbb{E}_{E(a),b,\epsilon,t}[\|\epsilon - \epsilon_\theta(a_{lt}, t, \mathcal{Z}_\theta(b))\|_2^2] \tag{6.3}$$

ControlNet: ControlNet is designed to control the diffusion models to enable them to perform a specific downstream task. It uses an input control map that provides an opportunity to manipulate the generated output. ControlNet, in its standard form, supports control maps with different conditions, such as edge maps, scribbles, segmentation maps, pose, etc. It preserves the weights of the SD by making a locked copy of it. Simultaneously, it uses a trainable copy with task-specific conditional control for a downstream task. These two copies are connected via 1×1 zero convolution layers with both bias and weight initialized as zero. The convolutional weights of these layers gradually optimize starting from zero, which gives the benefit of no extra added noise with faster training at the same time. Let parameters of locked copy be denoted as ϕ and those of trainable copy as ϕ_c. If zero convolution operation is $\mathcal{C}(.;.)$ which uses two instances of parameters $\{\phi_{c1}, \phi_{c2}\}$, then combining it into the ControlNet network blocks $\mathcal{H}(.;.)$ could be represented as:

$$d_{\mathfrak{c}} = \mathcal{H}(e, \phi) + \mathcal{C}(\mathcal{H}(e + \mathcal{C}(c, \phi_{\mathfrak{c}1}); \phi_{\mathfrak{c}}); \phi_{\mathfrak{c}1}) \tag{6.4}$$

where $d_{\mathfrak{c}}$ is the output and e is the input feature map. The overall objective after including the downstream task can be modified as shown below:

$$L_{CN} := \mathbb{E}_{E(a),b,b',\epsilon,t}[\|\epsilon - \epsilon_\theta(a_{lt}, t, \mathcal{Z}_\theta(b), \mathcal{Z}_\theta(b'))\|_2^2] \tag{6.5}$$

where $\mathcal{Z}_\theta(b')$ is the intermediate representation of the task-specific conditional text prompt.

6.2 ControlPolypNet

6.2.1 Methodology

Overview

The objective of the proposed method is to generate polyp frames to increase the sample count for training and enhance the performance of deep learning models. Given a set of images with two subsets, polyp/positive (P) and non-polyp/negative (N), our goal is to utilize images in N to expand the subset P. This is achieved by transforming images $N = \{n_1, n_2, ..., n_s\}$ into $P' = \{p'_i | p'_i$ is similar in distribution to $p_j\}$, where $p_j \in P$. Moreover, during this transformation, polyp shape, location, and size are user-configurable and integrating P' with P should diversify the overall set. This signifies that the synthetic polyp set P' should be diverse and possess qualities similar to real images in set P.

Architectural Details

The architectural details of our model *ControlPolypNet* are depicted in Fig. 6.2. It consists of three main parts: (a) an SD U-Net architecture loaded with pre-trained weights of SD v1-5, (b) ControlNet, and (c) YOLOv8, a detector pre-trained on the polyp images. The decoder part of the SD U-Net is kept unlocked, and only its encoder part is left locked during the complete training process. This unlocking is done to obtain better performance on medical imaging tasks like ours, as the initial weights are more inclined toward general images. Instead of adopting standard control map options presented by ControlNet, we tailored the input condition map to fit the necessary requirements.

We utilized the negative colonoscopy frames, which are relatively easily accessible in sufficiently large amounts. We overlapped these frames with random custom masks to obtain N', which are the regions targeted for polyp generation to obtain P'. To make the model learn the mapping $N' \rightarrow P'$, we prepared our training set such that initially, it learns $M \rightarrow P$, where M is obtained by overlapping P with its binary mask ground truth. By providing P as the target image and M as the source image (control image), the model learns the mapping $M \rightarrow P$. While learning this mapping, the model learns the complex patterns in data, and when given a random mask over non-polyp image n'_i, it transforms it into p'_i when given the text prompt "polyp". This mapping allows the usage of custom masks with controllable polyp positions and shapes. Also, this reduces the probability of obtaining unwanted structures or noise in the background/endoluminal scene.

When given a polyp image p_i, the standard diffusion process progressively adds noise

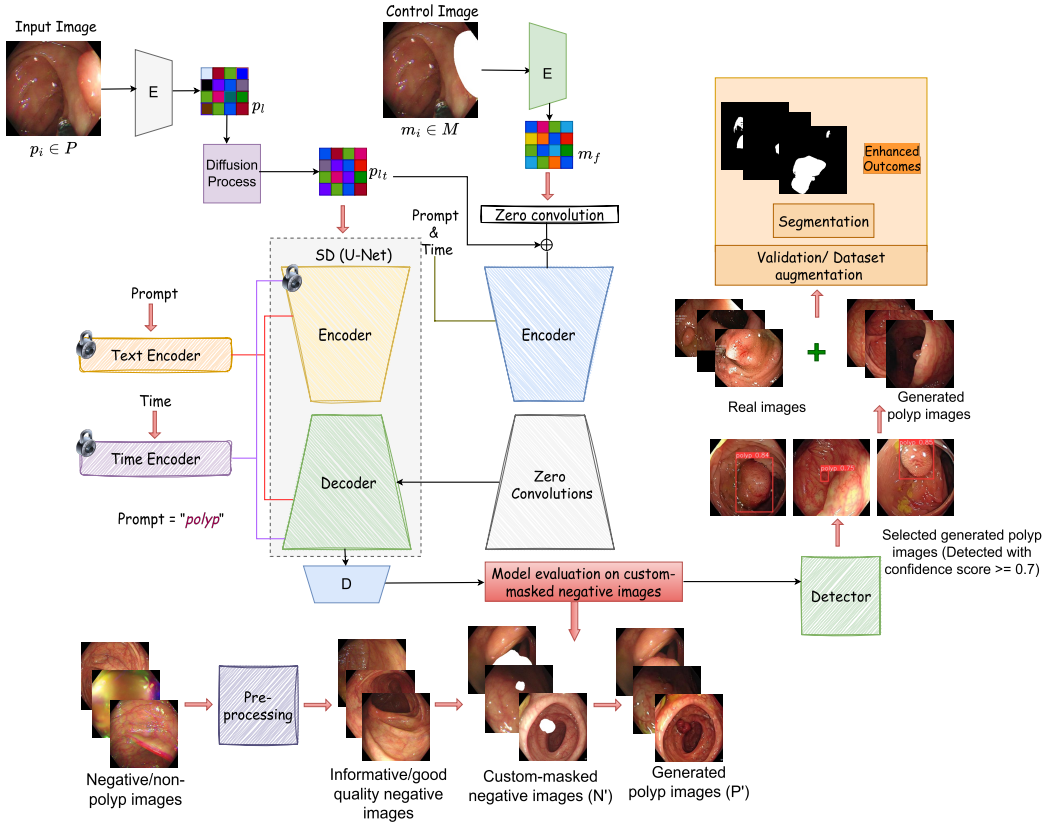


Figure 6.2: The proposed framework uses custom-masked images as control input with a "polyp" text prompt. The below pipeline shows the pre-processing of negative samples to remove uninformative frames and then using custom masks to generate polyps during the evaluation phase of ControlPolypNet. The complete process helps to enhance the segmentation task by providing the possibility of data augmentation, which ultimately validates the significance of synthetic polyp images.

to the image in its latent representation p_l to obtain a noisy version p_{lt}. This input is combined with conditions in the form of mask-overlapped image $m_i \in M$ and text prompt b, i.e., "polyp". m_i is further converted into an intermediate representation m_f by performing encoding on m_i to match the input size of SD. The objective of *ControlPolypNet* can be defined as:

$$L_{CPN} := \mathbb{E}_{E(p_i),b,m_f,\epsilon,t}[\|\epsilon - \epsilon_\theta(p_{lt}, t, \mathcal{Z}_\theta(b), m_f)\|_2^2] \tag{6.6}$$

The proposed input control ensures that the other endoluminal scene remains intact, which could be beneficial to capturing and differentiating polyp regions during downstream tasks. As stated in [16], considering some regions from background aids in improving classification results. This outcome could be attributed to polyps exhibiting a distinct color

and texture, setting them apart from the normal mucosal regions. Unwanted noise and irrelevant objects in the generated outputs create unrealistic data that could deviate the segmentation models from the intended tasks. Therefore, we utilized the negative frames instead of relying on the standard binary masks. However, these negative frames can have some artifacts, as general colonoscopy videos are prone to motion blur, interlacing, ghost colors, etc. Hence, we used an approach given in Chapter 1 to eliminate such uninformative negative frames before using them for the translation.

Pathological Validation Setup: Although generative models are now common in the medical imaging domain, various studies [196, 197] show that they are liable to generate unrealistic medical conditions or structures. As pathological patterns are significantly crucial, we performed an elimination step instead of directly integrating them into the segmentation task training. This elimination step validates the presence of lesion-characterizing features in the synthetic images and simultaneously prepares a clinically valid set of images appropriate for data augmentation. We integrated a polyp detector, YOLOv8 [198], in the proposed framework for this process. This detector is pre-trained on polyp images with a confidence score set in the range of 0.7 and 0.8 for inference. This integration helps choose only valid, visually appealing frames with lesion-characterizing features. We used these selected generated polyp frames to augment the training set for the segmentation task.

6.2.2 Experiments and Results

Dataset Details and Training Settings

We used three publicly available datasets, namely, SUN Database [95] (49,136 polyp frames and 109,554 non-polyp frames), CVC-ClinicDB [10] (612 polyp images) and Kvasir-SEG [9] (1000 polyp images), to validate the performance of our proposed framework. The segmentation ground truth of the SUN Database, released in the form of SUN-SEG [199], is also used. The SUN Database and SUN-SEG are used in the training of ControlPolypNet, whereas CVC-ClinicDB and Kvasir-SEG are used to validate generated image quality in the downstream tasks of polyp segmentation.

During *ControlPolypNet* training, we used 38,284 polyp images in the train set, and the rest were used for validation purposes. To translate non-polyp images into polyp images, we custom-masked 10,000 negative images obtained after pre-processing five non-polyp video sequence cases with the informative/uninformative frame detector given by [200]. The official split of CVC-ClinicDB and Kvasir-SEG is used. The implementation is done using PyTorch and PyTorch lightning frameworks. *ControlPolypNet* and downstream tasks

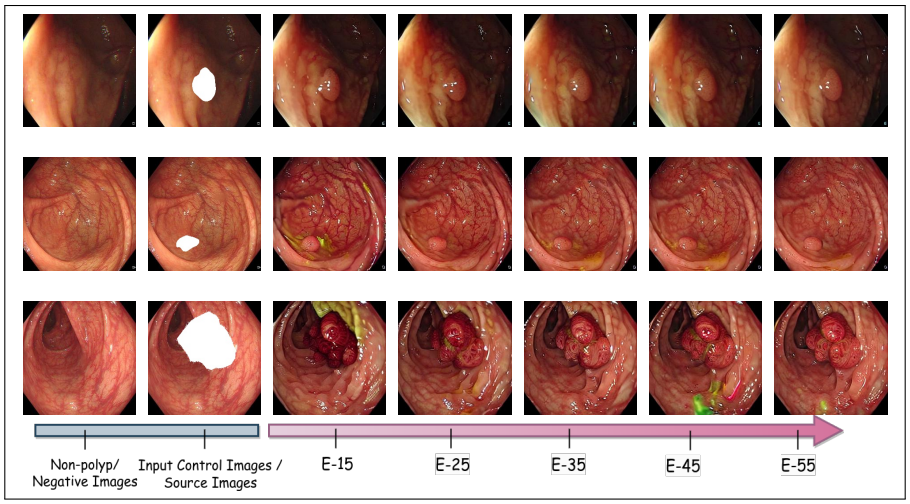


Figure 6.3: Epoch-wise sample images along with their corresponding negative images and input control images (custom-masked negative samples). E stands for epochs.

Table 6.1: Quantitative comparison of synthetic polyp images with different sets of real images over different epochs. **Bold** values represent the 'best' metrics score, and E, P, NP stand for 'epoch', 'polyp', and 'non-polyp'. ↓ and ↑ denote 'lower is best' and 'higher is best', respectively.

Metrics	Trend	Comparsion (with)	E-15	E-25	E-35	E-45	E-55
FID	↓	Real P images	104.52	106.70	102.46	99.35	**94.07**
		Real NP images	92.10	93.77	91.16	89.22	**82.95**
PSNR	↑	Masked NP images	67.70	67.22	67.66	67.57	**68.39**
SSIM	↑	Masked NP images	0.9987	0.9984	0.9986	0.9986	**0.9988**

training are executed using NVIDIA A100 and NVIDIA Titan-Xp GPU, respectively. *ControlPolypNet* is trained for 55 epochs with a batch size of 32 and a learning rate of $2e^{-6}$.

Evaluation Metrics

The quality of the generated images is accessed using three metrics: FID, PSNR, and SSIM. FID quantifies the quality of synthetic data for realism and diversity. PSNR is focused on the reconstruction quality of images, and SSIM quantifies the similarity between two images. Additionally, we used task-specific metrics for segmentation which include precision, recall, F1-score and JI. The JI determines the overlap between the ground truth and prediction masks.

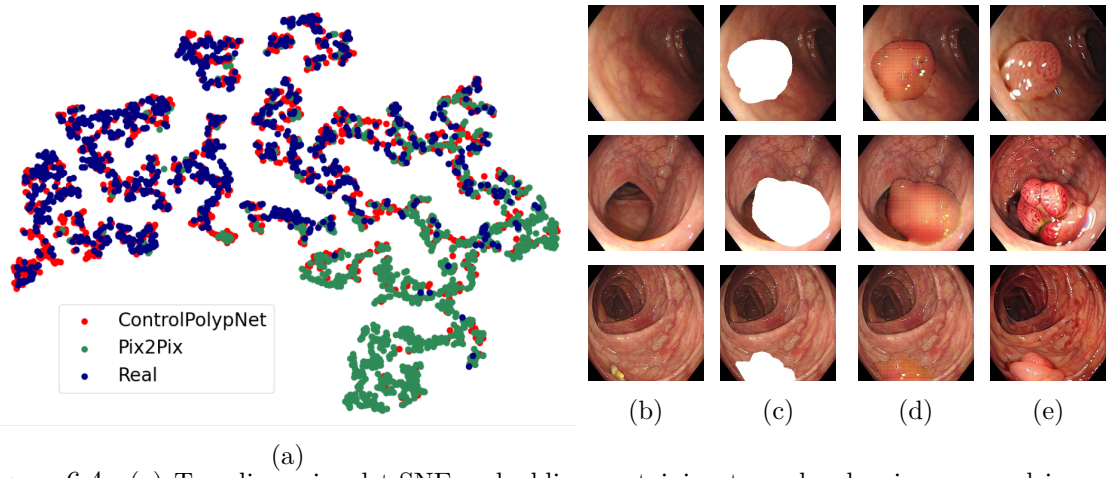


Figure 6.4: (a) Two-dimensional t-SNE embedding pertaining to real polyp images, and images generated by Pix2Pix and ControlPolypNet, (b)-(e) show negative images, masked negative images, synthetic images obtained using Pix2Pix and ControlPolypNet, respectively.

Performance Evaluation

We evaluated our model on different epochs and examined the quality of the generated images using the quality assessment metrics. The related results are shown in Table 6.1. While using FID, we considered two comparison scenarios: synthetic vs. real polyp images and synthetic vs. real non-polyp images. As expected, the latter case presented a better score because the related non-polyp images are translated into synthetic polyp with background details substantially preserved. It can be observed that the quality of images in both cases gradually increased with the epoch counts. Due to the high computational requirements of diffusion models, we considered training till the point where visually appealing results were obtained. We further explored the structure and information-preserving ability of our approach using PSNR and SSIM. We masked the generated images' polyp region for this assessment and compared them with the masked non-polyp images. The results show that the quality of the endoluminal scene is satisfactorily preserved and is improved with the increasing epochs.

Besides quantitative outcomes, we observed the qualitative results, shown in Fig. 6.3. In the initial epochs, especially in epoch 15, the image details are not precisely generated and are obstructed by artifacts. Moreover, the color transfer ability from the input control image to synthetic images is higher in the later epochs. The randomness in polyp color and close mapping of the polyp shape and its location with the custom mask demonstrates our approach's potential to achieve data diversity and successful control over synthetic polyp shape and location. Although the results demonstrate the scope of improvement in color-preservation ability, structural-preservation outcomes are impressive. Further, we compared

Table 6.2: Performance of the U-Net [66], ColonSegNet [62], and TransNetR[202] models on the downstream task of polyp segmentation. RI stands for Real Images. The best results are highlighted in **bold** and the second best are underlined.

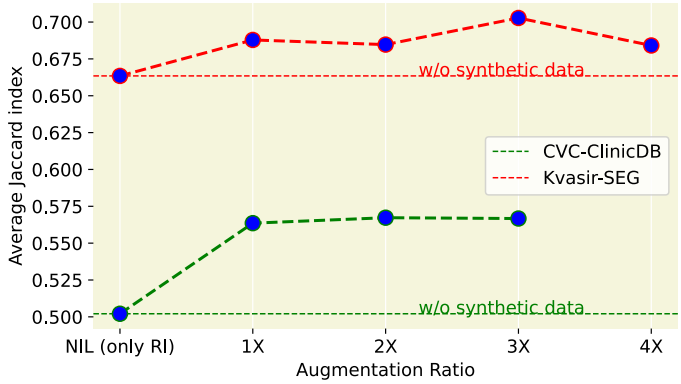
Dataset: CVC-ClinicDB												
Training sample count (x = 490)	**U-Net**				**ColonSegNet**				**TransNetR**			
	Jaccard	**Recall**	**Precision**	**F1-score**	**Jaccard**	**Recall**	**Precision**	**F1-score**	**Jaccard**	**Recall**	**Precision**	**F1-score**
RI (x)	0.4682	0.5211	0.8509	0.5523	0.3429	0.3834	0.8256	0.4424	0.6952	0.7431	0.9399	0.7737
RI + Random Rotation (x+x)	0.4748	0.5244	0.8909	0.5568	0.4352	0.4859	0.8161	0.5312	0.7015	0.7450	0.9468	0.7805
RI + Gaussian Blur (x+x)	0.4447	0.4809	0.8705	0.5215	0.3467	0.3779	0.8291	0.4453	0.6960	0.7433	0.9357	0.7762
RI + Vertical Flip (x+x)	0.4589	0.5027	**0.9218**	0.5354	0.3666	0.3976	0.8412	0.4585	0.6675	0.7094	0.9283	0.7442
RI + Horizontal Flip (x+x)	0.4348	0.5138	0.8447	0.5198	0.4296	0.4696	0.8991	0.5080	0.6991	0.7581	0.9279	0.7823
RI + Elastic Transformation (x+x)	0.4296	0.4696	0.8991	0.5080	0.3867	0.4275	0.8019	0.4874	0.5907	0.6197	0.9439	0.6691
RI + Pix2Pix Synthetic Images (x+x)	0.4493	0.4964	0.7917	0.5474	0.3872	0.4019	0.8343	0.4661	0.7076	0.7406	0.9469	0.7872
RI + ControlPolypNet Synthetic Images (x+x)	0.5356	0.5781	0.9096	0.6232	0.4360	0.4831	0.8211	0.5359	0.7191	0.7731	0.9366	0.7967
RI + Pix2Pix Synthetic Images (x+2x)	0.3363	0.4323	0.6736	0.4429	0.4196	0.4465	0.7680	0.5065	0.6953	0.7299	**0.9570**	0.7719
RI + ControlPolypNet Synthetic Images (x+2x)	0.5424	0.6390	0.8292	0.6365	0.4272	0.4828	0.7782	0.5267	0.7322	0.7837	0.9366	0.8113
RI + Pix2Pix Synthetic Images (x+3x)	0.4763	0.4975	0.8752	0.5570	0.4283	0.4531	0.8683	0.5192	0.6875	0.7174	0.9571	0.7599
RI + ControlPolypNet Synthetic Images (x+3x)	0.5375	0.5802	0.8660	0.6149	0.4726	0.5432	0.8093	0.5760	0.6900	0.7287	0.9505	0.7628
RI + 5 aug. (x+5x)	0.5518	0.6252	0.9002	0.6353	0.4928	0.5307	0.8623	0.5855	0.7214	0.7639	0.9426	0.7963
RI + 5 aug. + ControlPolypNet Synthetic Images (x+5x+2x)	**0.6298**	**0.7132**	0.8900	**0.7160**	**0.5928**	**0.6308**	**0.9167**	**0.6874**	**0.7486**	**0.7968**	0.9365	**0.8198**
Dataset: Kvasir-SEG												
Training sample count (x= 880)	**U-Net**				**ColonSegNet**				**TransNetR**			
	Jaccard	**Recall**	**Precision**	**F1-score**	**Jaccard**	**Recall**	**Precision**	**F1-score**	**Jaccard**	**Recall**	**Precision**	**F1-score**
RI (x)	0.6668	0.7796	0.8420	0.7508	0.5782	0.7148	0.7610	0.6869	0.7454	0.8273	0.9058	0.8267
RI + Random Rotation (x+x)	0.6852	0.7679	**0.8702**	0.7669	0.6143	0.7280	0.8045	0.7148	0.7469	0.8289	0.9005	0.8298
RI + Gaussian Blur (x+x)	0.6704	0.7736	0.8521	0.7563	0.5677	0.7116	0.7705	0.6793	0.7596	0.8426	0.8956	0.8399
RI + Vertical Flip (x+x)	0.6738	0.7693	0.8614	0.7580	0.6129	0.7504	0.7965	0.7184	0.7749	0.8552	0.8946	0.8501
RI + Horizontal Flip (x+x)	0.6837	0.7984	0.8390	0.7743	0.6039	0.7202	0.8105	0.7115	0.7629	0.8357	0.9120	0.8370
RI + Elastic Transformation (x+x)	0.6667	0.7996	0.8239	0.7538	0.6163	0.7399	0.8088	0.7208	0.7369	0.8265	0.8806	0.8160
RI + Pix2Pix Synthetic Images (x+x)	0.6550	0.7516	0.8353	0.7357	0.5757	0.6976	0.7920	0.6824	0.7659	0.8482	0.9020	0.8425
RI + ControlPolypNet Synthetic Images (x+x)	0.6795	0.8032	0.8498	0.7688	0.6262	0.7532	0.8098	0.7345	0.7579	0.8497	0.8801	0.8373
RI + Pix2Pix Synthetic Images (x+2x)	0.6127	0.7258	0.8103	0.7060	0.5820	0.7123	0.7783	0.6887	0.7651	0.8539	0.8984	0.8439
RI + ControlPolypNet Synthetic Images (x+2x)	0.6680	**0.8465**	0.7971	0.7640	0.6065	0.7508	0.7913	0.7209	0.7797	0.8665	0.9010	0.8523
RI + Pix2Pix Synthetic Images (x+3x)	0.6580	0.7624	0.8440	0.7441	0.6048	0.7353	0.7916	0.7113	0.7747	0.8524	0.9109	0.8497
RI + ControlPolypNet Synthetic Images (x+3x)	0.6997	0.8331	0.8464	0.7879	0.6326	0.7603	0.8121	0.7379	0.7760	**0.8677**	0.8938	0.8517
RI + Pix2Pix Synthetic Images (x+4x)	0.6720	0.7665	0.8633	0.7564	0.6021	0.7231	0.7961	0.6986	0.7346	0.8550	0.8441	0.8208
RI + ControlPolypNet Synthetic Images (x+4x)	0.6750	0.8126	0.8339	0.7651	0.6341	0.7835	0.7967	0.7440	0.7432	0.8139	0.9039	0.8245
RI + 5 aug. (x+5x)	0.7069	0.8131	0.8465	0.7912	0.6958	0.8086	0.8515	0.7907	**0.7960**	0.8518	**0.9366**	**0.8641**
RI + 5 aug. + ControlPolypNet Synthetic Images (x+5x+3x)	**0.7301**	0.8368	0.8657	**0.8153**	**0.7215**	**0.8191**	**0.8638**	**0.8129**	0.7861	0.8622	0.9024	0.8584

the outcomes of *ControlPolypNet* with the synthetic images generated using Pix2Pix [201]. We selected Pix2Pix because it uses a mechanism to translate images from one domain to another, suitable for our objective to translate $N' \rightarrow P'$. A qualitative comparison is shown in Fig. 6.4 where the images in Fig. 6.4(b)-(e) clearly show that although Pix2Pix retained the polyp location and shape, more realistic polyp images with texture were generated by *ControlPolypNet* while preserving the shape and location. However, compared to our model, Pix2Pix was better at retaining the original colors of background regions. Additionally, we generated a t-SNE plot (shown in Fig. 6.4(a)) using a DenseNet-201 that is trained to differentiate polyp and non-polyp images [200]. Plotting feature embeddings of real polyp images and synthetic polyp images generated by Pix2Pix and *ControlPolypNet* clearly depict the closeness of our model's outcomes with real images. Contrarily, generated images of Pix2Pix barely show any overlap with real data.

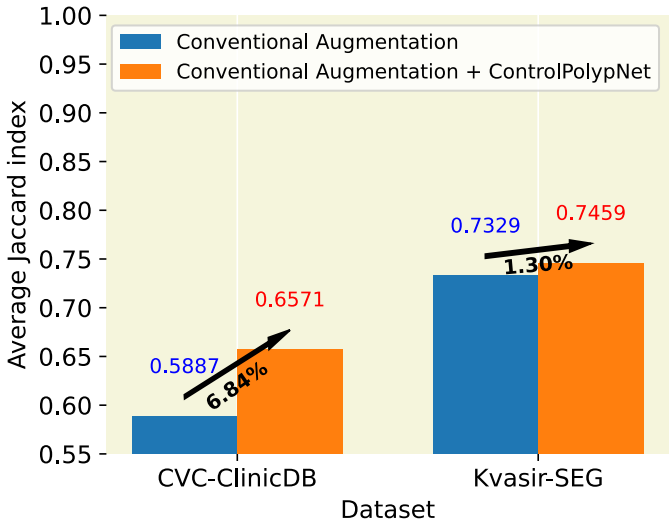
Clinical Significance Validation and Downstream Tasks Evaluation

The clinical significance validation step employs a detector, as discussed in Section 6.2.1. The synthetic images that were detected by YOLOv8 with confidence scores in the range of 0.7 and 0.8 are used to augment the dataset of the downstream task. This approach of using

synthetic images to augment medical data provides two-fold benefits: a) It validates the quality and clinical significance of the generated data, and b) It allows enhancing segmentation outcomes. We carried out experiments with different proportions of synthetic images and five general augmentations, namely, random rotation, Gaussian blur, elastic transformations and horizontal and vertical flips. The associated results are shown in Table 6.2. We used three state-of-the-art polyp segmentation models, U-Net [66], ColonSegNet [62], and TransNetR [202] to experiment with different data augmentation combinations.



(a) Impact of different data augmentation ratios. The polyp segmentation performance achieves a significant increase with small ratios, and then with increasing ratios, the improvement is either minimal or absent.



(b) Comparison between the two average JI obtained using conventional augmentations with and without images generated by ControlPolypNet.

Figure 6.5: Comparative analysis of average JI obtained after computing mean over all three segmentation models in different scenarios.

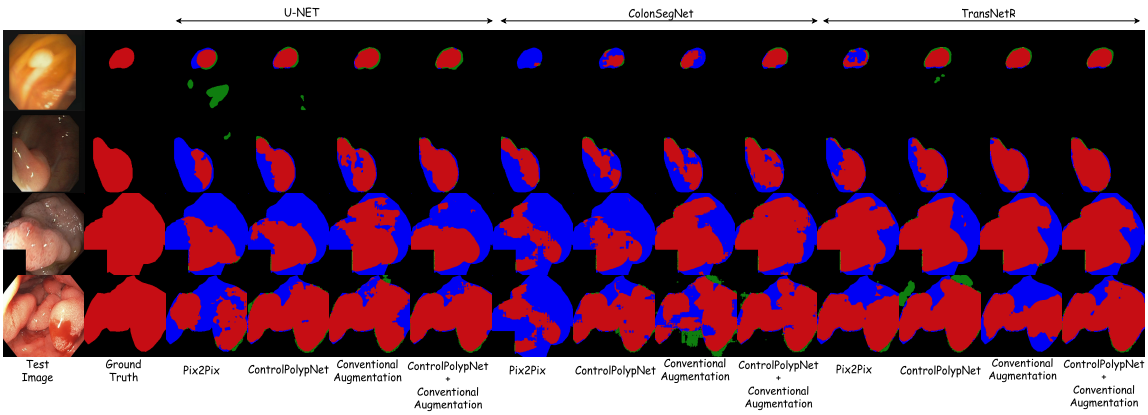


Figure 6.6: Qualitative results of polyp segmentation outcomes. The figure illustrates that in most cases when ControlPolypNet's output is combined with conventional augmentation techniques, it predicts masks closer to ground truth. Also, the mask obtained using ControlPolypNet's generated images performs better than synthetic images obtained using Pix2Pix.

Table 6.3: Quality assessment of generated images using Pix2Pix and ControlPolypNet. This assessment is conducted using U-Net [66], ColonSegNet [62], and TransNetR[202] models trained on real images. The best results are highlighted in **bold**.

Training Dataset	Generation Method	U-Net				ColonSegNet				TransNetR			
		Jaccard	Recall	Precision	F1-score	Jaccard	Recall	Precision	F1-score	Jaccard	Recall	Precision	F1-score
CVC-ClinicDB	Pix2Pix	0.2048	0.6943	0.2562	0.3054	0.2323	0.7129	0.2934	0.3414	0.4517	0.6510	**0.6219**	0.5662
	ControlPolypNet	**0.2613**	**0.7792**	**0.3088**	**0.3802**	**0.2633**	**0.7353**	**0.3328**	**0.3876**	**0.4761**	**0.7729**	0.5991	**0.6149**
Kvasir-SEG	Pix2Pix	0.5802	0.6994	**0.7450**	0.6597	**0.4778**	0.7508	**0.5814**	**0.5814**	0.6037	0.6414	**0.9109**	0.6657
	ControlPolypNet	**0.6285**	**0.8128**	0.7394	**0.7362**	0.4039	**0.7973**	0.4842	0.5354	**0.7580**	**0.8537**	0.8749	**0.8454**

During augmentation, we increased the ratio of synthetic images as a multiple of x, where x is the original training set size. It can be observed that adding synthetic images in x proportion performs comparable to adding a single conventional augmentation. We gradually increased synthetic images in ix proportion, where i={1,2,3,4}. The results show that the polyp segmentation performance achieves a significant increase with small ratios, and then, with increasing ratios, the improvement is either minimal or absent. The same can be inferred from Fig. 6.5(a). The value of i is incremented until the metrics values start to decrease. The proportion *i*x that performs the best is combined further with conventional augmentations. The outcomes from this integration show that synthetic images complement conventional augmentation techniques as the average performance increased compared to cases where only conventional augmentations were used. Additionally, we compared *ControlPolypNet* with Pix2Pix using the same proportion of their generated data for augmentation. An average Jaccard index over all the different proportions (x, 2x, 3x or 4x) is 5.61% and 2.3% higher using *ControlPolypNet* compared to Pix2Pix on CVC-ClinicDB and Kvasir-SEG, respectively. This increase can be observed in Fig. 6.5(b). Moreover, the individual performance with different data proportions and models has reported en-

hanced performance using our augmentation approach. It is noteworthy that even though the synthetic images are generated using a different larger dataset, they are performing effectively on a small out-of-distribution dataset. This observation supports both quality and diverse information possessed by the generated images. Adopting traditional augmentation techniques is limited by the actual size of the dataset as they can only be scaled up by its multiple. Also, this scaling up produces redundant information in some form. Contrarily, adding our diverse set of synthetic images can complement this information and is independent of real dataset size, thus providing enhanced segmentation outcomes. These results are supported by some qualitative outcomes, shown in Fig. 6.6. It can be observed that, in most cases, combining conventional techniques with *ControlPolypNet*'s synthetic data provides results closer to the ground truth. We further tested the synthetic images obtained using *ControlPolypNet* and Pix2Pix using the three segmentation models (trained using only real data). The results shown in Table 6.3 signify that our proposed approach generates more realistic images with polyp-specific characteristics.

Although our proposed approach provides an opportunity to obtain customized polyp images using negative images, some lingering gaps still need to be addressed. Controlling polyp location, size, and shape enables us to obtain diverse polyp images. However, control over colors remains unexplored. In medical images, color is one of the criteria considered for domain shift issues, as color variations across inter-hospital and inter-patient data bring performance drops in segmentation. Control over colonoscopy image color can expand the possibility of domain transfer and even enhance segmentation outcomes.

6.3 PathoPolyp-Diff

6.3.1 Methodology

An overview of our approach is illustrated in Fig. 6.7. Our model *PathoPolyp-Diff* utilizes dual-stage training. The two stages, *Step-I* and *Step-II*, aim at generating colonoscopy images with diverse polyp types in different imaging modalities. They perform complementary tasks, and the difference between the two lies in their training process. The *Step-I* network distils knowledge into the *Step-II* model in the form of a large set of features that enables it to differentiate between polyp and non-polyp characteristics and further helps to generate images for cross-class labels. Complementary to it, the training process in *Step-II* allows the model to learn pathological details and different imaging modality-related patterns.

Step-I: The Step-I uses a pre-trained SD v1-4[1] model and further fine-tunes it with some text conditions to generate desired colonoscopy images. This stage is focused on developing a model that learns the basic features to differentiate polyp and non-polyp characteristics. In the process, text prompts are used as the conditioning mechanism to control the model output. These text prompts comprise the embeddings pertaining to the strings presented at the first to third levels in Fig. 6.8, starting from the top. During the fine-tuning of the model, a relatively large-scale dataset is used with polyp/non-polyp classes wherein the polyps have an additional annotation of low-quality/artifacts (uninformative) and good-quality/clear (informative). This process allows the model to generate polyp and non-polyp images with specific quality criteria.

Step-II: In this stage, the pre-trained model of Step-I is used with the first block locked. The other blocks are further fine-tuned on our desired text conditioning. These conditions are shown in the fourth and fifth levels of Fig. 6.8. For a successful implementation of cross-class label learning, we undersampled the non-polyp set and informative set of polyps and included them in the training iterations of Step-II. Presenting these undersampled images allows the model to retain the features pertaining to Step-I without undergoing overfitting with the new dataset.

6.3.2 Results

Dataset Details and Training Settings

We used two publicly available datasets, namely, SUN Database [95, 203] and ISIT-UMR Colonoscopy Dataset [96], to evaluate the performance of our proposed model. SUN Database comprises 49,136 polyp frames and 109,554 non-polyp frames, where polyp frames have additional annotations for informative/uninformative classes (annotated in Chapter 1). The dataset consists of 31% uninformative polyp frames and is used to train our model in step-I. ISIT-UMR Colonoscopy Dataset is an NBI and WLI video dataset and consists of labels for hyperplastic, adenoma and sessile classes. We used the first two of these classes to train our model in step-II. The same dataset is used for validation purposes during classification. We converted 40 adenoma and 21 hyperplastic video streams of each NBI and WLI modalities into frames. 29 adenoma and 15 hyperplastic videos of both modalities are used for training *PathoPolyp-Diff*. During the classification task, 15 videos and 6 videos (2281 frames of adenoma class, 1618 frames of hyperplastic class) of each pathology and modality are used in the train set and test set, respectively.

[1]https://github.com/CompVis/stable-diffusion

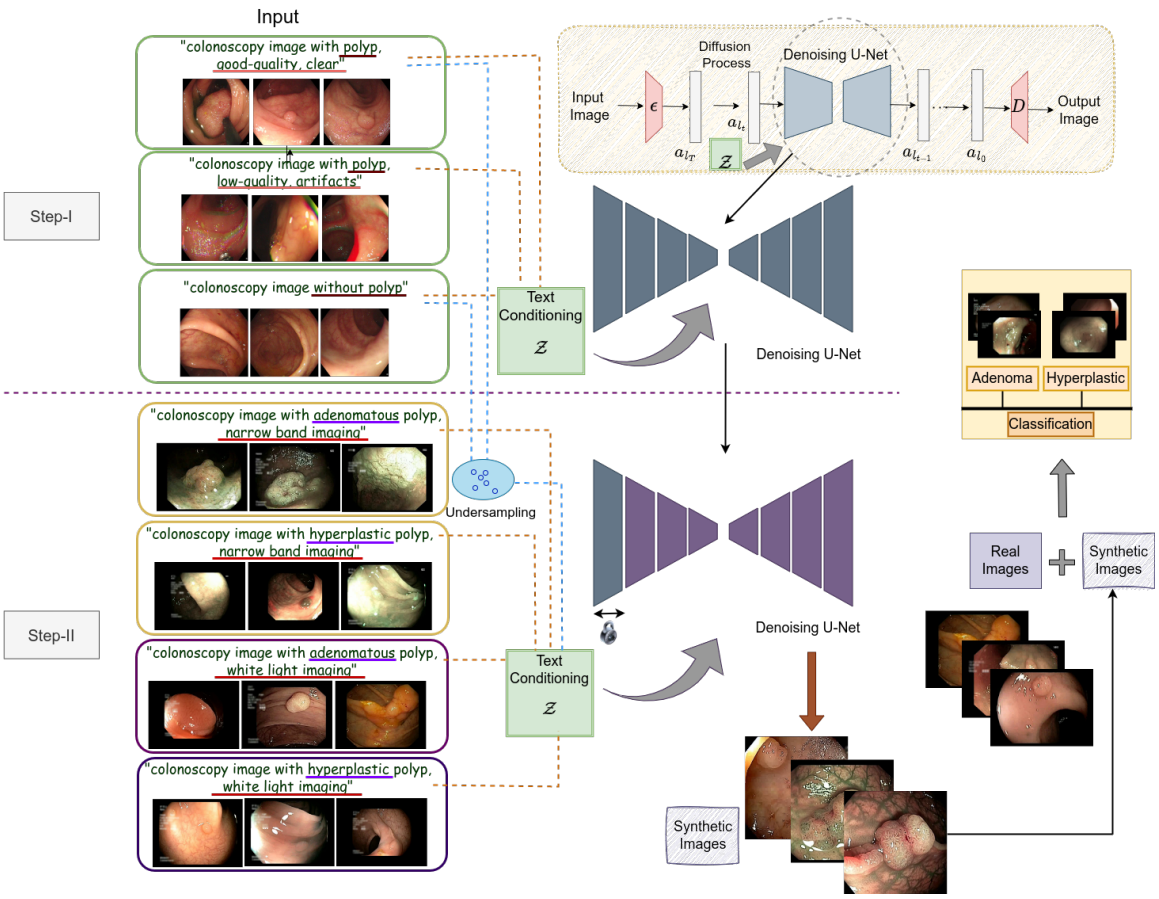


Figure 6.7: Overview of the proposed framework. It consists of two steps and uses various text conditioning to control the generation process. In Step-II, some undersampled data from Step-I is used for a smoother learning process. Also, the first block of U-Net is kept locked in the second step. The performance of the proposed model is validated using a classification process which uses a combination of real and synthetic images in different proportions.

During *PathoPolyp-Diff* training in Step-I, a pre-trained Stable Diffusion v1-4 model was loaded which was further fine-tuned on the SUN Database with a learning rate, batch size and resolution set to 1e-06, 15, and 512, respectively. The best model chosen for subsequent fine-tuning in Step-II was identified after 8,000 iterations. In step-II, ISIT-UMR Colonoscopy Dataset is used for training along with the same learning rate and one block of U-Net locked. All the implementations are carried out using the PyTorch framework and experiments are executed on NVIDIA A100 and NVIDIA Titan-Xp GPU for *PathoPolyp-Diff* and classification tasks, respectively.

Evaluation Metrics: In this work, we adopted some standard metrics used for classification. It includes precision, recall, F1-score, and BA. The first three are commonly used to evaluate any classification model. The last metric, i.e. BA, is generally used when an

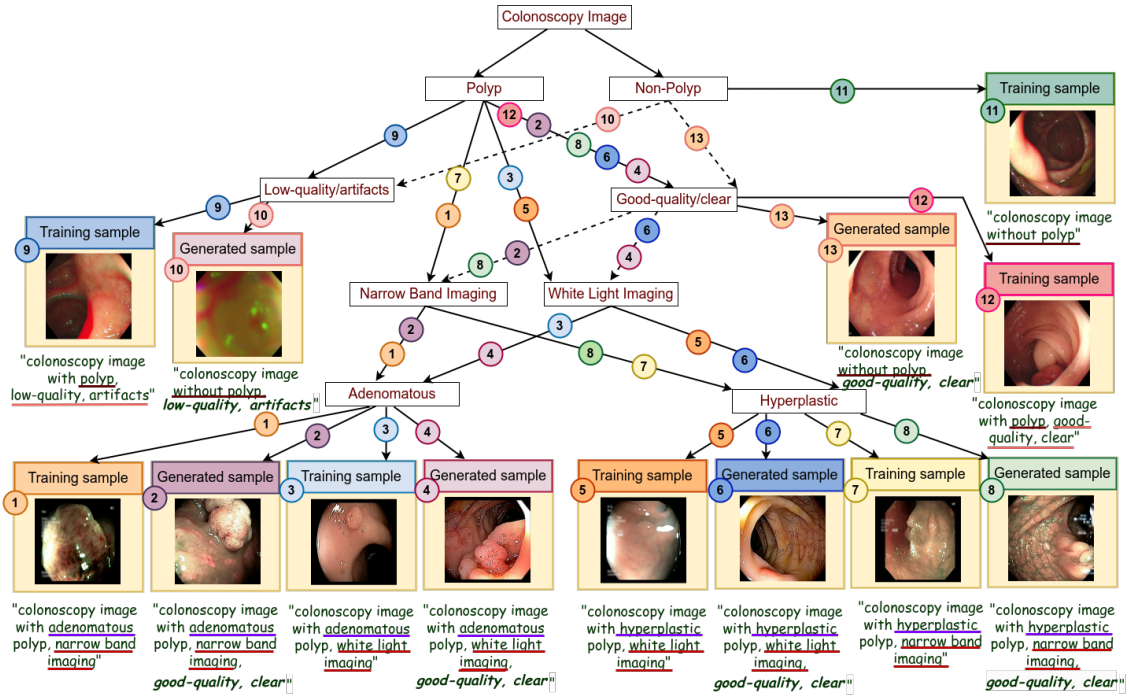


Figure 6.8: Flowchart depicting the different combinations of text prompt and cross-class labels used to generate images. The solid arrows denote the labels already present in the dataset, whereas the dashed arrows represent the labels learnt from other classes (cross-class labels). Each number on a solid/dashed line represents the combination of strings used to form tokens for text prompts used in training/inference. For instance, following number *'8'*, we obtain the text prompt *"colonoscopy image with a hyperplastic polyp, narrow band imaging, good quality, clear"*, where *"good quality, clear"* are part of indirectly inferred tokens and other are already present in the training annotations.

imbalance is encountered in data distribution, and the objective focuses on both minority and majority classes with equal importance during evaluation. This scenario aligns with our case, where we consider both adenomatous and hyperplastic classes to be equally important, as the clinical treatments depend on the diagnosed pathology class. The BA can be defined as an arithmetic mean of specificity (true negative rate) and sensitivity (true positive rate). In addition, we used Kernel Inception Index (KID) [204] for quality assessment of the generated images. It quantifies the dissimilarity between the generated and real data distributions.

Model Performance

Step-I: To evaluate the performance of our model in generating polyp and non-polyp images and to select the best model for the subsequent training process, we used four assessment metrics, namely, KID, precision, recall, and F1-score. Initially, we trained the

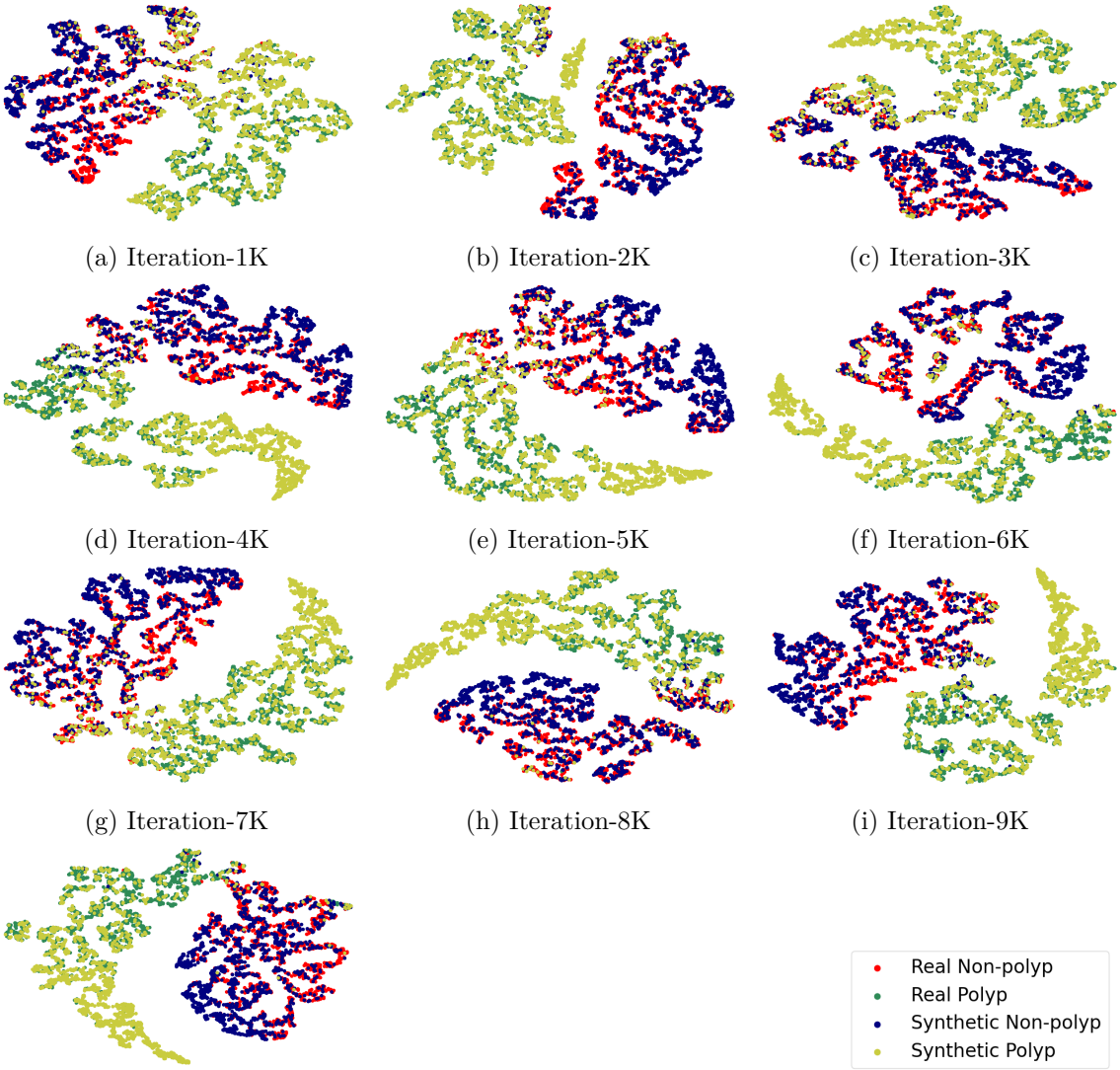
Table 6.4: Iteration-wise quality assessment of generated images in Step-I. This assessment is done using KID (similarity with real images), precision, recall, and F1-score (polyp/non-polyp characterizing features). ↓ and ↑ denote 'lower is best' and 'higher is best', respectively.

Prompt	Metrics	Behavior	Iterations									
			1K	2K	3K	4K	5K	6K	7K	8K	9K	10K
colonoscopy image with polyp, good-quality, clear	Precision	↑	0.8671	0.9085	0.9617	0.9720	0.9850	0.9858	**0.9961**	0.9859	0.9761	0.9899
	Recall	↑	0.9567	**0.9933**	0.8367	0.9267	0.8733	0.9267	0.8467	0.9300	0.9533	0.9800
	F1-score	↑	0.9097	0.9490	0.8948	0.9488	0.9258	0.9553	0.9153	0.9571	0.9646	**0.9849**
	KID	↓	**0.045**	0.059	0.055	0.075	0.053	0.071	0.048	0.051	0.049	0.064
colonoscopy image without polyp	KID	↓	**0.027**	0.041	0.037	0.034	0.067	0.051	0.051	0.045	0.055	0.057
	Average KID	↓	**0.036**	0.05	0.046	0.0545	0.06	0.061	0.0495	0.048	0.052	0.0605

model for 10,000 iterations and selected the model after every 1,000 iterations for testing. The corresponding assessment results are given in Table 6.4. As our main focus is to obtain good-quality polyp images, we generated and validated images with substring *"good-quality, clear"*. Additionally, we used a DenseNet-201 [113], pre-trained to detect polyp frames (proposed in Chapter 3), to validate the polyp-characterizing features existing in the generated polyp images. Therefore, we tested generated polyp and non-polyp images using DenseNet-201. It can be observed that the lowest (also the best) average KID of 0.036 is obtained at $1,000^{th}$ iteration; however, it reported a low F1-score. Similar outcomes are achieved with the next lowest KID. Contrarily, the highest F1-score in $10,000^{th}$ iteration is obtained with 0.0605 KID, the second lowest among all iterations.

To study the reasoning behind the contradictory results, we plotted t-SNE embeddings, shown in Fig. 6.9. It can be observed that in the initial iterations (1,000 to 5000), the polyp and non-polyp features of synthetic data are not entirely distinct; therefore, the associated F1-scores are relatively low. At the same time, they are finely overlapping with their real counterparts, therefore resulting in lower KID. To establish a trade-off between the KID and F1-score, we leveraged the visualization capabilities of t-SNE plots. Finally, the model at 8000^{th} iteration is selected as its KID score (in terms of all categories, i.e., polyp, non-polyp and average) is higher than the average score computed over each corresponding KID category. For instance, the average KID_{polyp} is calculated as $KID_{polyp_{1k}} + KID_{polyp_{2k}} + KID_{polyp_{3k}} ... + KID_{polyp_{10k}}$, which comes out to be 0.57. It can be noted that $KID_{polyp_{8k}} < KID_{polyp}$, and a similar observation can be identified in other categories. Moreover, the t-SNE plots signify that after 8000^{th} iteration, the synthetic polyp and non-polyp features start to deviate from the feature space, representing their real counterparts. As one class of synthetic features is still far apart from the other class of synthetic features, the F1-score is higher at the last iterations.

Impact of Negative Prompt: Negative prompt is an additional parameter which



(a) Iteration-1K (b) Iteration-2K (c) Iteration-3K

(d) Iteration-4K (e) Iteration-5K (f) Iteration-6K

(g) Iteration-7K (h) Iteration-8K (i) Iteration-9K

(j) Iteration-10K

Figure 6.9: Iteration-wise two-dimensional t-SNE embeddings to visualize the data points pertaining to synthetic and real polyp/non-polyp images.

guides the process of generating synthetic images not to include some specific objects or characteristics. It can assist in eliminating unwanted elements from the synthetic images. Hence, to further improve the quality of generated images, we evaluated our model on various negative prompts, including *"low-quality"*, *"blur"* and *"low-quality, blur"*. This approach is experimented with both polyp and non-polyp frames, and quality assessment is performed using the fast Fourier transform (FFT). The results shown in Fig. 6.10 present a significant improvement when a normal text prompt is combined with our specific quality-based negative prompt. The best combination is achieved when we club text prompt with *"blur, low-quality"*. Therefore, the subsequent experiments include this specific negative

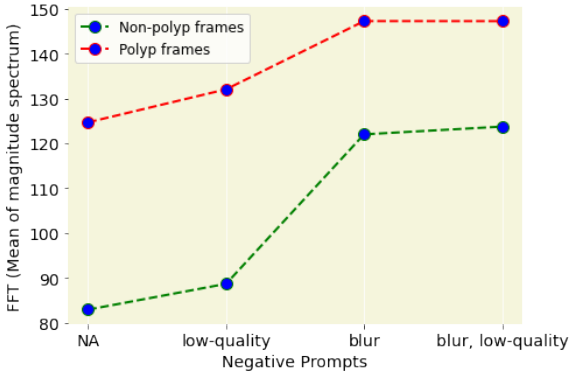


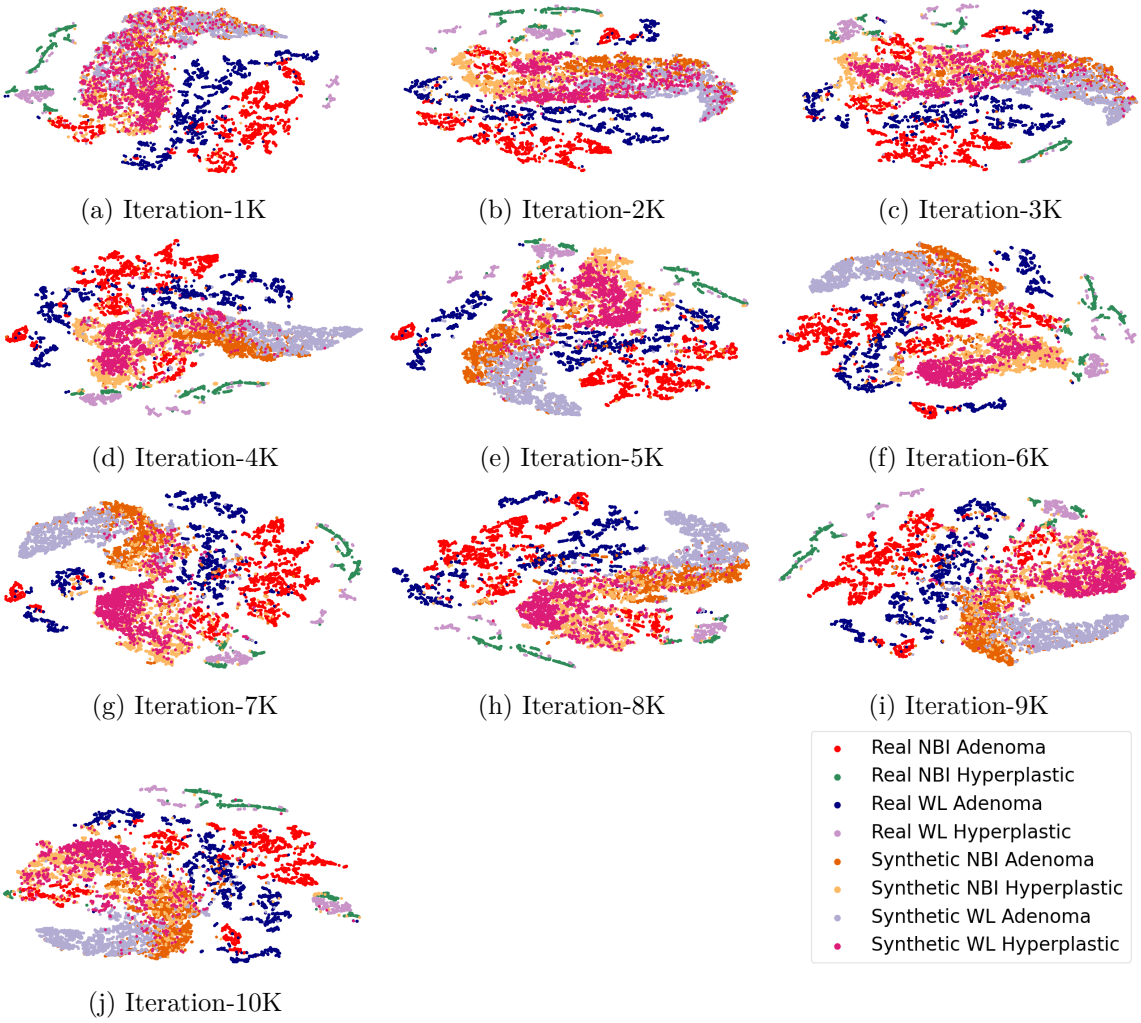
Figure 6.10: Quality assessment for validating the impact of negative prompt using FFT.

Table 6.5: Class-wise quality assessment of generated images after every 1000 iterations during Step-II. ↑ denotes 'higher is best'.

Class	Metrics	Behavior	Iterations									
			1K	2K	3K	4K	5K	6K	7K	8K	9K	10K
Adenoma	Precision	↑	0.50	0.54	0.57	0.65	0.69	**0.71**	0.65	0.59	0.61	0.58
	Recall	↑	**0.99**	0.98	**0.99**	**0.99**	0.98	0.97	0.98	**0.99**	**0.99**	**0.99**
	F1-score	↑	0.67	0.69	0.73	0.79	0.81	**0.82**	0.78	0.74	0.76	0.73
Hyperplastic	Precision	↑	0.75	0.90	**0.98**	0.97	0.97	0.96	0.95	0.96	**0.98**	0.95
	Recall	↑	0.02	0.15	0.26	0.48	0.55	**0.61**	0.48	0.33	0.37	0.29
	F1-score	↑	0.04	0.26	0.42	0.64	0.70	**0.74**	0.64	0.49	0.54	0.45

prompt to enhance quality, as it helps the model avoid generating blurred or low-quality images.

Step-II: To assess the quality of images generated in Step-II, specifically in terms of pathological characteristics, we performed binary classification. This binary classification is done for two different purposes: (a) To select the best model among different iterations (validation similar to Step-I) and (b) To verify how the synthetic images impact the classification performance if used for augmentation. In the first case, we used a pre-trained DenseNet-201 model intended to classify adenomatous and hyperplastic classes from WLI and NBI modalities. The synthetic images generated from every 1000^{th} epoch are evaluated using this model and the related results are shown in Table 6.5. Similar to Step-I, we plotted t-SNE feature embeddings in Step-II, to analyze the overlay and proximity of synthetic features to real images. As illustrated in Fig. 6.11, in the initial epochs, the pathology-specific features are not learnt, and hence, a significant overlap is observed among the generated images from different classes. It can also be inferred from the findings presented in Table 6.5, where DenseNet-201 exhibited significant challenges in effectively distinguishing between the two classes until 3000^{th} iteration. These outcomes are further supported by the confusion matrices in Fig. 6.12. These matrices demonstrate the biased shift of the model



(a) Iteration-1K (b) Iteration-2K (c) Iteration-3K

(d) Iteration-4K (e) Iteration-5K (f) Iteration-6K

(g) Iteration-7K (h) Iteration-8K (i) Iteration-9K

(j) Iteration-10K

Figure 6.11: Iteration-wise two-dimensional t-SNE embeddings to visualize the data points pertaining to synthetic and real adenomatous/hyperplastic images involving NBI/WLI imaging modalities.

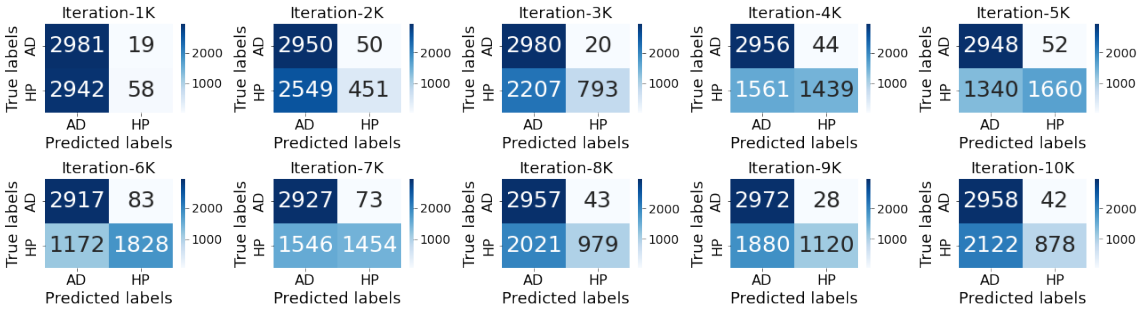


Figure 6.12: Confusion matrices to validate the iteration-wise performance of our model in generating adenomatous/hyperplastic polyp images with NBI/WLI imaging modalities.

towards the adenomatous class, which gradually improves with increasing epochs and, after some epochs, again shows the same biased performance. This trend is observed due to the

deviation of synthetic data from the expected pathological behavior as we train the model after a certain number of iterations. This analysis is based on the last few plots in Fig. 6.11. Considering the results in Table 6.5, Fig. 6.11, and Fig. 6.12, we selected the model at 6000^{th} iteration which reports the highest F1-score for both the classes (adenoma: 0.82, hyperplastic: 0.74).

In the second case (i.e. case (b)), we used the synthetic images to augment the real data. The effectiveness of using synthetic images is validated using a binary classifier, EfficientNet-B0 [184]. In addition to validating the synthetic data inclusion, we also compared the quality of synthetic images obtained using two different text prompts. Table 6.6 shows the associated frame-wise results with various data proportions. Starting with real data samples of 16, 32 and 64 per video, we subsequently increased the sample count by adding synthetic images in ix proportion, where $i = \{1, 2, 3\}$. This procedure is followed for '*A*' and '*B*' text prompts, which can be defined as *"colonoscopy image with* ***p__type*** *polyp,* ***p__mod****"* and *"colonoscopy image with* ***p__type*** *polyp,* ***p__mod****, good-quality, clear"*, respectively where **p__type** denote adenomatous/hyperplastic and **p__mod** denote NBI/WLI. It is noteworthy that the label *"good-quality, clear"* used in the text prompt '*B*' is not directly related to the training samples provided to *PathoPolyp-Diff* during Step-II; instead, they are learnt from a different dataset through cross-class label learning during Step-I. The comparative analysis of synthetic images pertaining to the two text prompts aims to assess the effectiveness of cross-class label learning. As shown in Table 6.6, with 16 real images per video, adding an equal number of synthetic images improves the frame-wise BA by 1.14% to 4.75% with NBI and WLI. A similar increase of 2.66% to 3.93% is reported when the ratio of synthetic data is doubled. With a further increase, i.e., when the proportion of synthetic data is three times, the results are enhanced by up to 7.91%. However, with 32 or 64 real images per video, the performance increase is relatively less significant and limited to about 3.5%. Therefore, it can be inferred that a relatively significant performance gain is achieved when a substantially small real dataset is merged with synthetic data. Further increasing the real dataset shows comparatively less improvement (without a monotonic trend) with a similar data augmentation approach.

Impact of Cross-class Label Learning: We further performed validation and analysis for the cross-label learning approach. It can be observed that in most cases, the BA and class-wise F1-scores using text prompt '*B*' are comparatively higher than that of text prompt A. Some notable improvements in BA include 0.6192 to 0.6614 (difference of +4.22%, 95% CI: 1.5%, 6.91%, p-value = 0.0068) using 64 real samples per video with an equal proportion

of synthetic data, 0.5561 to 0.6254 (difference of +6.93%, 95% CI: 5.01%, 8.85%, p-value < 0.0001) using 16 real samples per video with two times samples of synthetic data, and 0.6228 to 0.6983 (difference of +7.55%, 95% CI: 2.88%, 12.22%, p-value = 0.0058) using 16 real samples per video with three times samples of synthetic data. These outcomes signify that the quality of generated data can be improved with variations in the input text prompts. Moreover, the labels used in these text prompts can be indirectly inferred from other classes, thus reducing the requirements of annotated data for each scenario.

Table 6.6: Classification results using different proportions of real and synthetic images. Text prompt *A*and *B*stand for *"colonoscopy image with* ***p__type*** *polyp,* ***p__mod****"* and *"colonoscopy image with* ***p__type*** *polyp,* ***p__mod****, good-quality, clear"*, respectively, where **p__type** denote adenomatous/hyperplastic and **p__mod** denote NBI/WLI.

Imaging Modality	Training sample count (Real + Synthetic)		Text Prompt	Adenoma			Hyperplastic			Balanced Accuracy
				Precision	Recall	F1-score	Precision	Recall	F1-score	
NBI	x=16 images per video	x	-	0.8173±0.026	0.2897±0.035	0.4266±0.037	0.4757±0.009	**0.9079±0.020**	0.6242±0.007	0.5988±0.012
		x+x	*A*	0.7876±0.122	0.3194±0.111	0.4360±0.099	0.4689±0.006	0.8449±0.121	0.6005±0.026	0.5821±0.007
			B	0.8146±0.080	0.3295±0.050	0.4671±0.056	0.4853±0.023	0.8910±0.059	0.6281±0.031	0.6102±0.037
		x+2x	*A*	0.6596±0.020	**0.4131±0.025**	**0.5077±0.021**	0.4580±0.012	0.6991±0.029	0.5533±0.016	0.5561±0.015
			B	**0.8299±0.042**	0.3569±0.032	0.4977±0.027	**0.4965±0.007**	0.8939±0.040	**0.6382±0.012**	**0.6254±0.011**
		x+3x	*A*	0.7359±0.059	0.3529±0.055	0.4730±0.046	0.4712±0.014	0.8133±0.076	0.5959±0.028	0.5831±0.022
			B	0.8133±0.113	0.3524±0.047	0.4857±0.027	0.4867±0.010	0.8680±0.097	0.6227±0.033	0.6102±0.025
	x=32 images per video	x	-	0.8099±0.083	0.3320±0.036	0.4690±0.040	0.4840±0.020	0.8843±0.068	0.6253±0.032	0.6082±0.034
		x+x	*A*	0.7713±0.078	0.3970±0.073	0.5170±0.051	0.4909±0.007	0.8192±0.093	0.6124±0.025	0.6081±0.014
			B	0.7634±0.055	0.3980±0.015	0.5222±0.008	0.4917±0.011	0.8221±0.054	0.6151±0.023	0.6101±0.021
		x+2x	*A*	0.7562±0.029	0.4155±0.024	0.5358±0.022	0.4958±0.012	0.8101±0.033	0.6150±0.017	0.6128±0.017
			B	0.8240±0.066	**0.4160±0.041**	**0.5510±0.035**	**0.5137±0.018**	0.8700±0.064	0.6455±0.028	**0.6430±0.029**
		x+3x	*A*	0.7797±0.083	0.3841±0.036	0.5115±0.024	0.4905±0.013	0.8371±0.074	0.6180±0.028	0.6106±0.024
			B	**0.8464±0.090**	0.3805±0.043	0.5221±0.040	0.5060±0.020	**0.8948±0.068**	**0.6460±0.030**	0.6377±0.031
	x=64 images per video	x	-	**0.8889±0.035**	0.3698±0.022	0.5216±0.019	0.5124±0.005	**0.9338±0.025**	**0.6615±0.007**	0.6516±0.008
		x+x	*A*	0.7704±0.026	0.4129±0.032	0.5368±0.027	0.4995±0.011	0.8255±0.033	0.6221±0.014	0.6192±0.014
			B	0.8463±0.062	**0.4405±0.021**	**0.5780±0.011**	**0.5277±0.010**	0.8823±0.060	0.6601±0.024	**0.6614±0.022**
		x+2x	*A*	0.7964±0.031	0.4120±0.016	0.5426±0.011	0.5062±0.007	0.8501±0.034	0.6344±0.014	0.6310±0.013
			B	0.7818±0.056	0.4033±0.013	0.5314±0.014	0.4988±0.013	0.8381±0.051	0.6252±0.024	0.6207±0.023
		x+3x	*A*	0.8348±0.080	0.4086±0.012	0.5477±0.023	0.5132±0.021	0.8810±0.069	0.6483±0.035	0.6448±0.036
			B	0.7900±0.078	0.3922±0.028	0.5218±0.014	0.4959±0.013	0.8447±0.073	0.6245±0.029	0.6185±0.025
WLI	x=16 images per video	x	-	0.7856±0.064	0.5689±0.084	0.6583±0.072	0.4237±0.068	0.6695±0.096	0.5176±0.074	0.6192±0.073
		x+x	*A*	0.7869±0.057	0.6247±0.058	0.6937±0.031	0.4373±0.041	0.6268±0.129	0.5126±0.062	0.6257±0.051
			B	0.8563±0.039	0.5296±0.051	0.6521±0.031	0.4444±0.010	**0.8037±0.070**	0.5714±0.015	0.6667±0.013
		x+2x	*A*	0.7709±0.037	**0.7551±0.049**	**0.7616±0.028**	0.4944±0.054	0.5144±0.111	0.5013±0.073	0.6348±0.048
			B	0.8385±0.044	0.5635±0.113	0.6658±0.072	0.4504±0.028	0.7534±0.119	0.5593±0.026	0.6585±0.020
		x+3x	*A*	0.7652±0.017	0.7196±0.052	0.7407±0.028	0.4687±0.029	0.5260±0.066	0.4937±0.035	0.6228±0.023
			B	**0.8621±0.053**	0.6302±0.156	0.7154±0.099	**0.5079±0.079**	0.7665±0.130	**0.6010±0.040**	**0.6983±0.039**
	x=32 images per video	x	-	0.8460±0.033	0.6760±0.089	0.7475±0.046	0.5188±0.047	0.7291±0.089	**0.6018±0.019**	**0.7026±0.016**
		x+x	*A*	0.7983±0.030	0.7695±0.089	0.7801±0.034	0.5493±0.050	0.5751±0.126	0.5514±0.046	0.6723±0.024
			B	0.8328±0.034	0.7009±0.023	0.7605±0.020	0.5199±0.017	0.6952±0.080	0.5939±0.039	0.6981±0.032
		x+2x	*A*	0.7868±0.024	0.7011±0.093	0.7395±0.059	0.4920±0.075	0.5957±0.051	0.5351±0.047	0.6484±0.044
			B	0.8454±0.040	0.5983±0.092	0.6963±0.057	0.4728±0.037	0.7591±0.097	0.5793±0.032	0.6787±0.032
		x+3x	*A*	0.7960±0.015	**0.7799±0.089**	**0.7857±0.045**	**0.5628±0.079**	0.5714±0.068	0.5602±0.029	0.6756±0.025
			B	**0.8542±0.056**	0.5664±0.186	0.6626±0.113	0.4701±0.072	**0.7675±0.157**	0.5696±0.020	0.6670±0.023
	x=64 images per video	x	-	0.8007±0.053	0.5690±0.149	0.6574±0.120	0.4432±0.084	0.7023±0.095	0.5389±0.073	0.6357±0.075
		x+x	*A*	0.7657±0.035	**0.7225±0.066**	**0.7408±0.019**	**0.4634±0.018**	0.5171±0.139	0.4824±0.073	0.6198±0.038
			B	0.8445±0.029	0.4436±0.062	0.5792±0.052	0.4096±0.019	**0.8229±0.050**	**0.5463±0.019**	0.6333±0.022
		x+2x	*A*	0.7767±0.025	0.6606±0.100	0.7107±0.060	0.4570±0.054	0.5928±0.083	0.5112±0.037	0.6267±0.035
			B	0.8183±0.062	0.4888±0.120	0.6016±0.086	0.4086±0.027	0.7499±0.143	0.5247±0.042	0.6194±0.036
		x+3x	*A*	0.7559±0.011	0.6762±0.023	0.7135±0.012	0.4343±0.013	0.5320±0.038	0.4778±0.020	0.6041±0.013
			B	**0.8230±0.043**	0.5547±0.075	0.6589±0.044	0.4362±0.021	0.7361±0.113	0.5451±0.042	**0.6454±0.033**

Video-wise Analysis with Statistical Significance Test

A comparative study has been conducted on patient-wise results. Although the training experiments are performed on the frame level, the inference is computed on both the frame and video levels. A majority voting scheme has been adopted for such computations. For instance, a video is labelled as the class 'adenomatous' if the majority of frames (> mean number of total video frames) are predicted as adenomatous. These results are provided

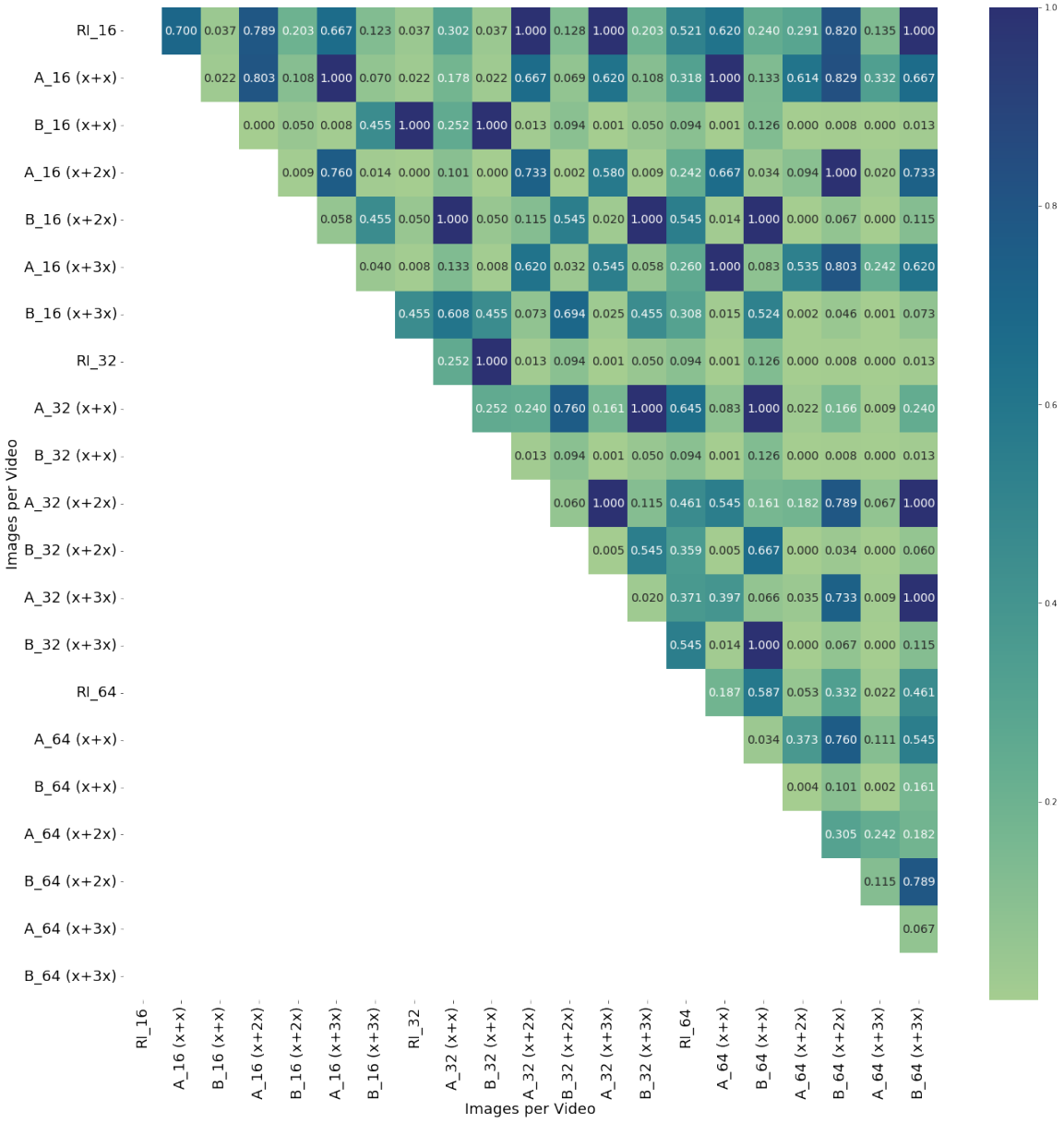


Figure 6.13: p-value obtained using two-tailed t-test for statistical significance analysis of video-wise (WLI) outcomes. The values are rounded off to 3 decimal places. The label names used in rows and columns can be read as "Text Prompt_Sample Count per Video (Real Image Count + Synthetic Image Count)".

in the Table 6.7 and Table 6.8. Additionally, a statistical significance test is conducted using a two-tailed t-test, which signifies the significance of an increase or decrease in video-wise outcomes and reports if this change is insignificant. This test calculates the p-values between all possible combinations of the data proportions used in our work. The associated p-values are depicted in Fig. 6.13 and 6.15. Using the WLI modality, we observed that the best outcomes are obtained using 32 real images per video, 16 or 32 real images per video

Training sample count (Real + Synthetic)		Text Prompt	Balanced Accuracy
x=16 images per video	x	-	0.6333±0.126
	x+x	*A*	0.6±0.137
		B	**0.7833±0.046**
	x+2x	*A*	0.6167±0.046
		B	0.7167±0.046
	x+3x	*A*	0.6±0.109
		B	0.75±0.083
x=32 images per video	x	-	**0.7833±0.046**
	x+x	*A*	0.7167±0.119
		B	**0.7833±0.047**
	x+2x	*A*	0.6333±0.095
		B	0.7333±0.037
	x+3x	*A*	0.6333±0.046
		B	0.7167±0.046
x=64 images per video	x	-	0.6833±0.109
	x+x	*A*	0.6±0.070
		B	**0.7167±0.075**
	x+2x	*A*	0.5667±0.037
		B	0.6167±0.095
	x+3x	*A*	0.5333±0.046
		B	0.6333±0.095

Table 6.7: Video-wise results using WLI modality.

Training sample count (Real + Synthetic)		Text Prompt	Balanced Accuracy
x=16 images per video	x	-	0.6667
	x+x	*A*	0.7167±0.095
		B	0.75±0.059
	x+2x	*A*	0.6833±0.037
		B	0.7667±0.037
	x+3x	*A*	**0.7833±0.046**
		B	0.75
x=32 images per video	x	-	0.7333±0.070
	x+x	*A*	0.7333±0.070
		B	0.7667±0.037
	x+2x	*A*	**0.8±0.046**
		B	0.7833±0.046
	x+3x	*A*	0.7667±0.037
		B	0.7833±0.046
x=64 images per video	x	-	**0.8167±0.037**
	x+x	*A*	**0.8167±0.037**
		B	0.8±0.046
	x+2x	*A*	**0.8167±0.037**
		B	0.7667±0.037
	x+3x	*A*	**0.8167±0.037**
		B	0.7667±0.037

Table 6.8: Video-wise results using NBI modality.

combined with synthetic images generated from text prompt '*B*' (either in equal or in double proportion). The last two cases align with the second-best results in frame-level evaluation presented in Table 6.6. Moreover, considering the case with 16 real images per video and an equal proportion of synthetic samples, video-level analysis reports a significant improvement of 15% (p-value = 0.037). Similar to frame-level analysis, the video-level validations signify

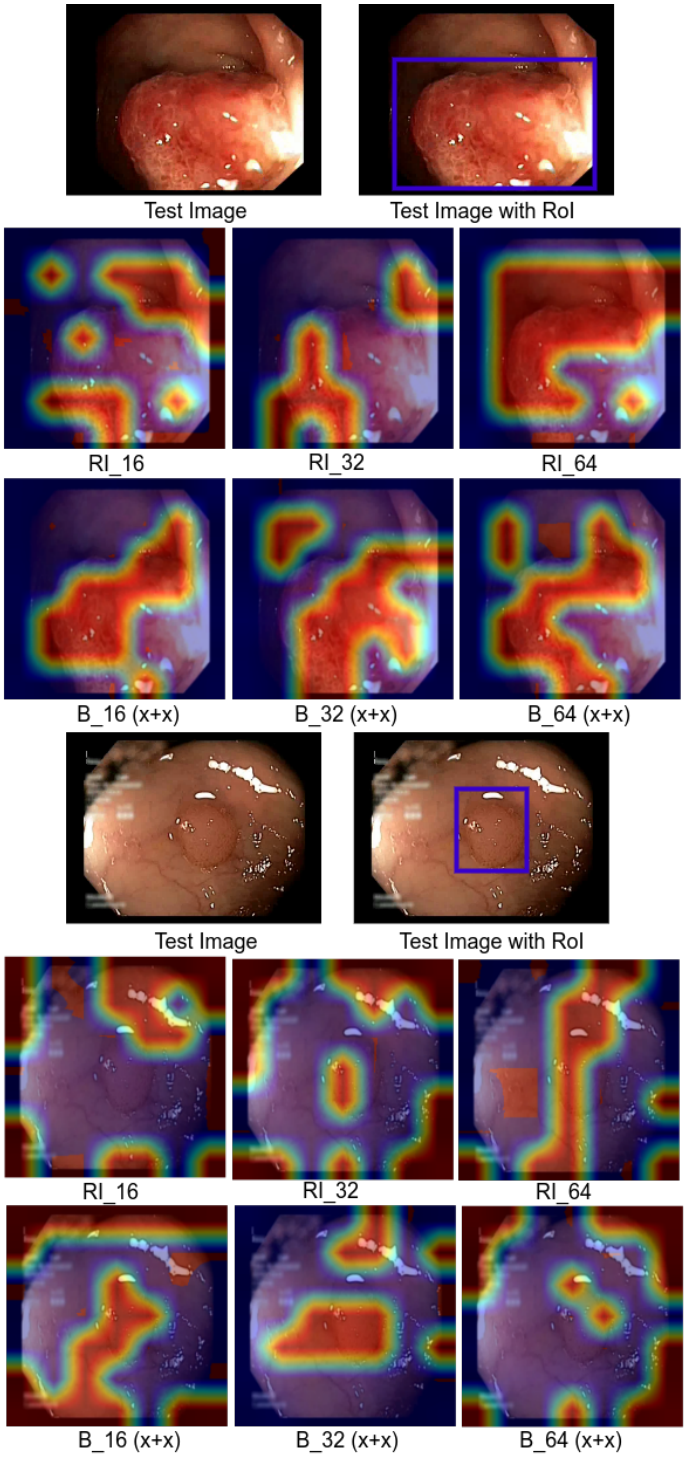


Figure 6.14: Heatmaps for the video-wise best outcomes obtained for both real and augmented data.

that the performance improvement with synthetic data reduces with increasing real sample count. Although with 64 real images per video, an increase of 3.34% is observed, this increase

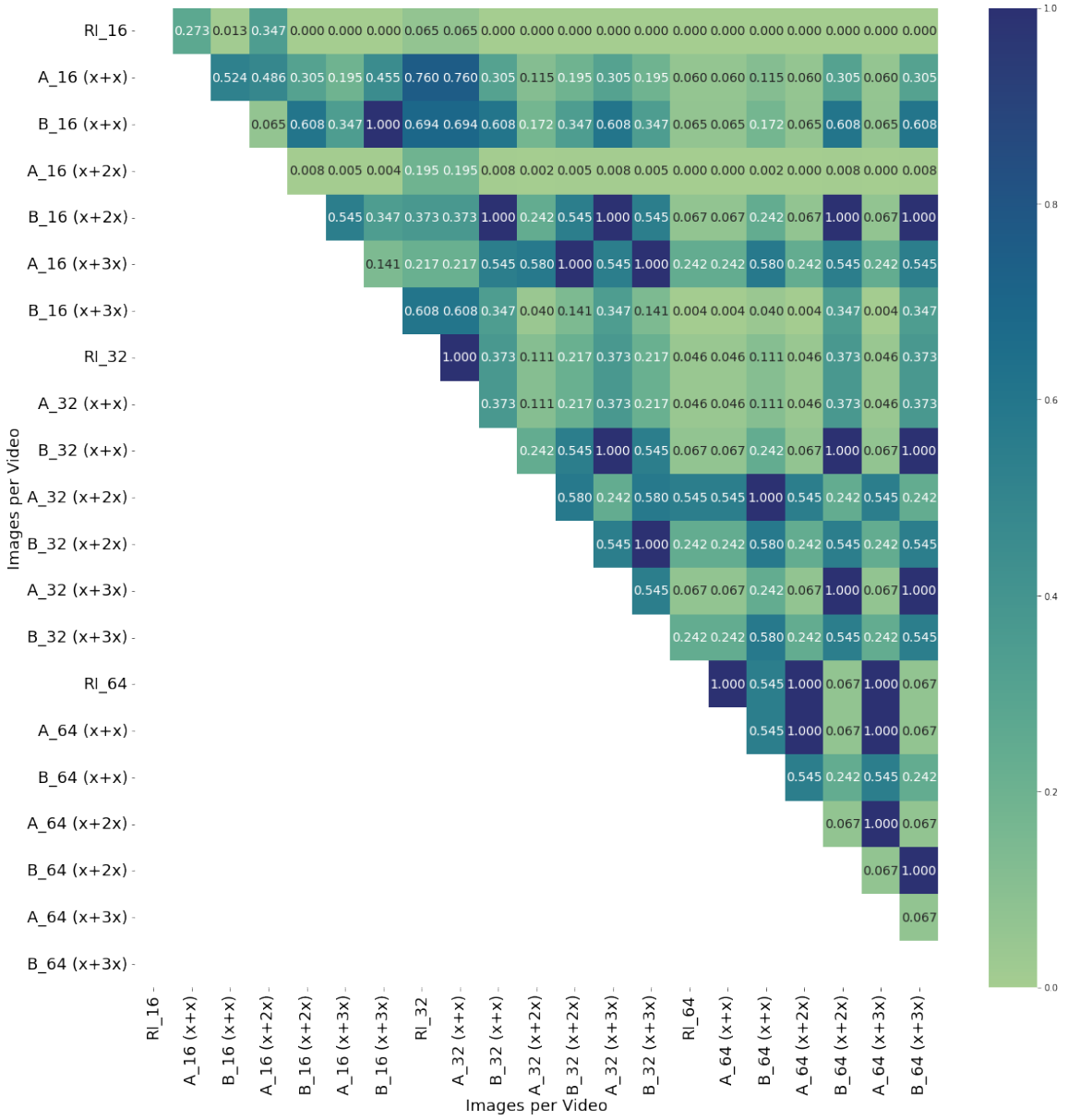


Figure 6.15: p-value obtained using two-tailed t-test for statistical significance analysis of video-wise (NBI) outcomes. The values are rounded off to 3 decimal places. The label names used in rows and columns can be read as "Text Prompt_Sample Count per Video (Real Image Count + Synthetic Image Count)".

is not statistically significant (p-value = 0.587). We further examined the performance difference between the two text prompts, '*A*' and '*B*', to evaluate the video-level cross-class label learning ability. Our analysis demonstrates that synthetic images generated using text prompt '*B*' are superior and statistically significant to those generated using text prompt '*A*'. This observation is supported by some of the notable improvements that include 0.6 to 0.7833 (difference of +18.33%, 95% CI: 3.43%, 33.23%, p-value = 0.022) using 16 real

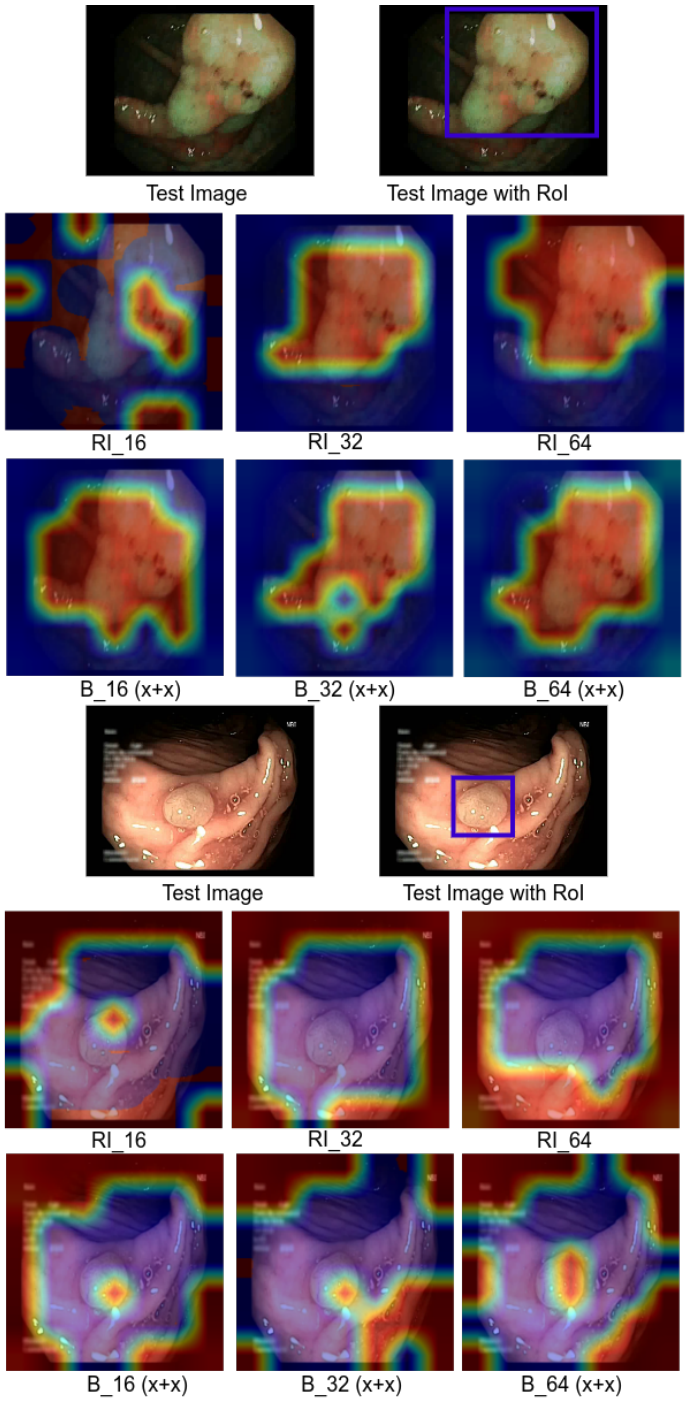


Figure 6.16: Heatmaps for the best outcomes obtained for both real and augmented data.

samples per video with an equal number of synthetic samples, 0.6 to 0.7167 (difference of +11.67%, 95% CI: 1.09%, 22.25%, p-value = 0.034) using 64 real samples per video with an equal number of synthetic samples, and 0.6333 to 0.7167 (difference of +8.34%, 95% CI: 1.63%, 15.05%, p-value = 0.020) using 32 real samples per video combined with three times

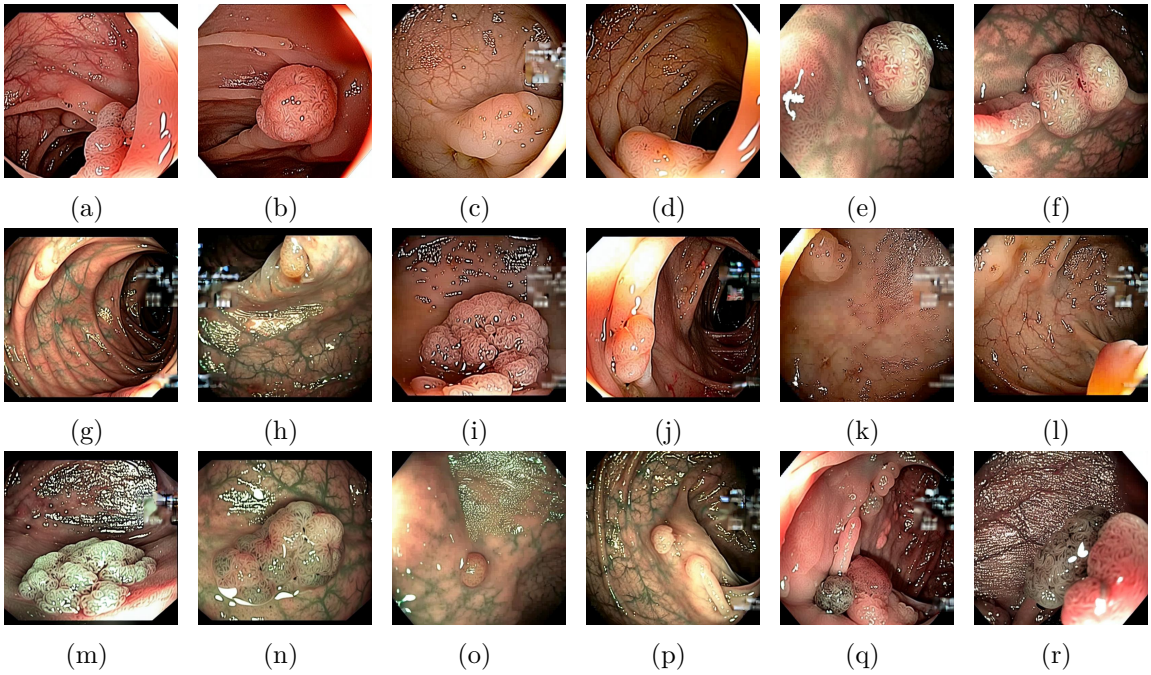


Figure 6.17: Sample generated images depicting (a)-(b) adenomatous polyp in WLI (using text prompt B) (c)-(d) hyperplastic polyp in WLI (using text prompt B), (e)-(f) adenomatous polyp in NBI (using text prompt B), (g)-(h) hyperplastic polyp in NBI (using text prompt B), (i)-(j) adenomatous polyp in WLI (using text prompt *A*), (k)-(l) hyperplastic polyp in WLI (using text prompt *A*), (m)-(n) adenomatous polyp in NBI (using text prompt *A*), (o)-(p) hyperplastic polyp in NBI (using text prompt *A*), and (q)-(r) shows some undesired images generated using text prompt *B*with NBI.

as many synthetic samples.

Using the NBI modality, the overall performance trend with data proportion is similar. Also, the comparison between the real data and a combination of real and synthetic data shows a similar trend as observed using the WLI modality. However, in the NBI modality, the synthetic data generated using text prompt '*A*' presented better outcomes than text prompt '*B*' in some cases. Although the best results are obtained using text prompt A, the average performance over all data proportions is the same for both text prompts. Moreover, it is noteworthy that the difference is not statistically significant in almost every case. Some such examples include the improvement from 0.7167 to 0.75 (difference of +3.33%, 95% CI: -8.2028% to 14.86%, p-value = 0.524) using 16 real images per video and an equal number of synthetic images, 0.7833 to 0.80 (difference of +1.67%, 95% CI: -5.04%, 8.38%, p-value = 0.580) using 32 real images per video and twice as many synthetic images, and 0.8167 to 0.7667 (difference of +5%, 95% CI: -0.40%, 10.40%, p-value = 0.067) using 64 real images per video and twice/thrice as many synthetic images. Despite such change in trend in video-wise analysis, it can be observed from Table 6.6 that on a frame-wise level, most of the cases

favored text prompt '*B*' over text prompt '*A*'. This inconsistent shift can be due to the fact that control over diffusion models is limited and also depends on the seed value. The original dataset used in Step-I with annotations based on the quality (good-quality, clear/ low-quality) comprises WLI images, whereas the WLI and NBI images used in Step-II lack such annotations. It was a relatively simple task for the model to generate a combination of WLI and good-quality data. On the contrary, the constrained control over the generated data occasionally resulted in blending WLI characteristics into some NBI images. This inconsistency emerged from the association between WLI and quality learned during Stage-I training. Random seed initialization and limited control over diffusion models resulted in some arbitrary outcomes with text prompt '*B*' in the case of NBI images. This justification is supported by the qualitative outcomes (shown in Fig. 6.17 (q)-(t)) discussed in the next section.

Qualitative Results and Interpretability through Visualization

In addition to the quantitative analysis, we examined the qualitative results and further studied the related heatmaps for visualization and interpretability. The heatmaps pertaining to the scenarios involving real images and also those related to the best model (with each 16, 32 or 64 cases) involving synthetic images are shown in Fig. 6.14 and Fig. 6.16. These heatmaps illustrate the region the classification model focuses on before providing the final prediction scores. It can be observed that the classifier learns the complex polyp-specific features better when trained using a more diverse set of polyp images obtained using PathoPolyp-Diff. However, the classifier's performance drops in identifying polyp features when the count of synthetic images increases. This decline in performance could be because synthetic images might carry noise and can not exactly replicate real image characteristics; thus, added noise could deviate the model after a certain limit.

Further, we analyzed the generated images for qualitative analysis. It can be observed that the synthetic images obtained using text prompt '*B*' are more visually appealing than those produced with text prompt A. The texture is more prominent and clear in images shown in Fig. 6.17(a) to Fig. 6.17(h) compared to those presented in Fig. 6.17(i) to Fig. 6.17(p). Moreover, in both scenarios, qualitatively, the generated images are close to real images in terms of structure, color and texture. The fundamental color criteria that differentiate NBI and WLI remain consistently evident in the images, making them easily distinguishable. However, as already discussed, text prompt '*B*' with NBI images fails for some samples, as can be inferred from Fig. 6.17(q) and Fig. 6.17(r). This failure is

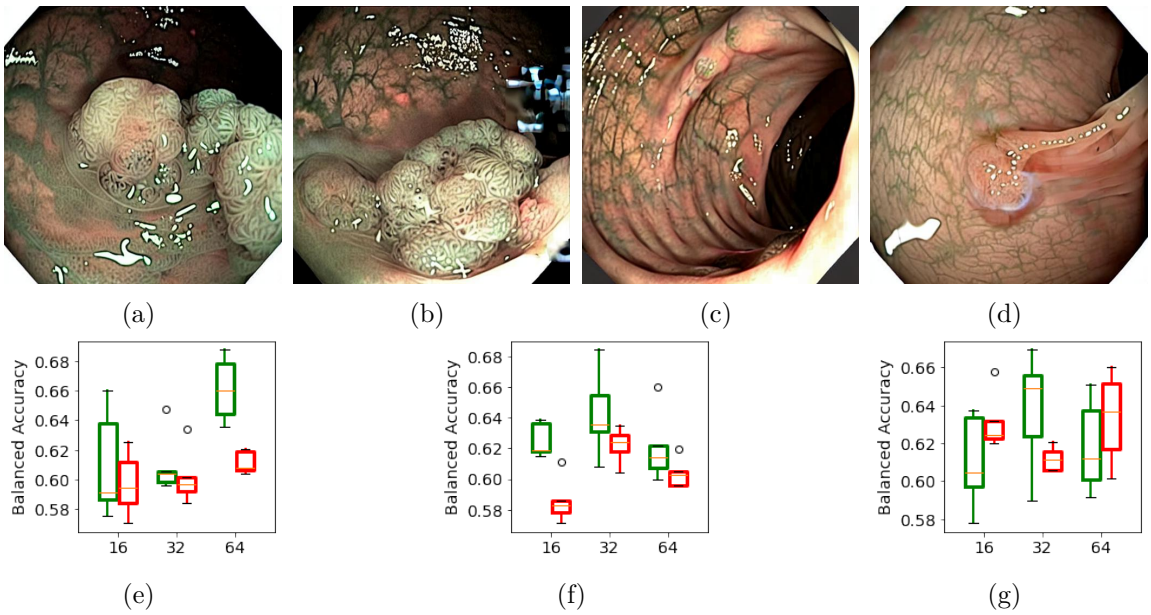


Figure 6.18: Sample images depicting (a)-(b) adenomatous polyps in NBI, and (c)-(d) hyperplastic polyps in NBI using weighted control mechanism. The boxplots (e) to (g) demonstrate the comparison between text prompt B with and without a weighted control mechanism when synthetic images are added in equal proportion or twice or thrice in proportion to the real images, respectively. The former and latter text prompts are denoted by green and red color, respectively.

attributed to the limited control over the image generation and dependency on seed initialization. Note that the pathology-focused data used in Step-II has both WLI and NBI images, but none of the images have annotations based on quality. Still, the model learnt the relation between quality and WLI easily because of the relation developed earlier in Step-I. Therefore, in Step-II, our model more precisely established the relation between WLI, quality, and pathology. Due to the development of this direct relationship, the model sometimes recollects WLI characteristics when it assigns more weightage to the tokens *"good-quality, clear"* in the given text prompt *"colonoscopy image with adenomatous polyp, narrow band imaging, good-quality, clear"* or *"colonoscopy image with hyperplastic polyp, narrow band imaging, good-quality, clear"*. Consequently, some samples present a combination of WLI and NBI images or complete WLI images.

To overcome this issue, we assigned more weight to the tokens representing *"colonoscopy image with* ***p__type*** *polyp, narrow band imaging"* where **p__type** can be hyperplastic/adenomatous. This modification resulted in better qualitative outcomes, as depicted in Fig. 6.18(a) to Fig. 6.18(d). However, the related quantitative outcomes are reduced in most of the cases covered by Fig. 6.18(e) to Fig. 6.18(g). One possible reason for such a decline in results could be due to a lack of pathology-specific features when images are

generated using a weighted approach. Such outcomes signify that slight modifications in the text prompt could impact synthetic images' characteristics and visual properties.

6.4 Chapter Summary

In this chapter, we propose two different diffusion based techniques to generate synthetic images. In the first technique, we proposed a stable diffusion based framework, ControlPolypNet, to generate polyp frames utilizing non-polyp frames. We showed that the polyp generation process can be customized, and different controls can be used to get more fine-grained data. The generated frames also capture pathological features with visually impressive results and help enhance the downstream tasks of polyp segmentation. A detector is introduced to ensure the retention of pathological features in our proposed framework. Using our approach, an average increase of 6.84% and 1.3% (JI) over three models is achieved on CVC-ClinicDB and Kvasir-SEG, respectively.

In the second technique, we developed a novel diffusion based model, PathoPolyp-Diff, which generates a realistic-looking diverse set of polyp images in a two-stage process. This set of polyps covers multiple categories, including pathology (adenomatous/hyperplastic), imaging modality (NBI/WLI), and quality (informative/uninformative). We validated the pathological content of the synthetic images by using the generated images to augment the real images of a public dataset. This augmented dataset is then used for the downstream task of binary classification on adenomatous/hyperplastic polyps. We compared our augmentation approach with the baseline (only real data) using different proportions of the dataset. The best result reported an increase of 7.91% (0.6983 ± 0.039 vs. 0.6192 ± 0.073). With a similar comparison approach (using different data proportions), we also examined the synthetic images generated using different variations of text prompts. It is observed that the text prompts formulated using the cross-class label concept outperformed those without such labels in most of the cases (for both NBI and WLI). In addition to frame-level analysis, we conducted video-level investigations. The associated quantitative results are supported by a statistical significance test (two-tailed t-test) and heatmaps. At video-level analysis, a statistically significant difference can be observed in favor of the cross-class label concept. Although the text prompts without cross-class labels achieved the best outcomes for NBI cases, the overall performance of both text prompts was similar, and the difference was not statistically significant. In addition, we explored the concept of weighted text prompts and presented both qualitative outcomes and quantitative analysis through box plots. We

also provide a roadmap for the research community to build upon our work, extending the synthetic polyp dataset and experimenting with text prompts to enhance overall outcomes.

7

Conclusions and Future Directions

This thesis explores avenues to address several challenges that may arise during the deployment of CAD systems in clinical settings. For this purpose, our research examines various phases involved in an automated system. We investigate the quality of the colonoscopy video frames and demonstrate their effect on clinical outcomes. In this context, we proposed solutions for extracting good-quality and clinically informative data for accurate diagnosis and effective treatment. Besides the quality assessment, this thesis addresses the critical issue of limited data availability by releasing a multi-class GI dataset. Moreover, we proposed techniques to generate synthetic data to mitigate the challenges associated with data scarcity and ethical aspects of the medical domain. Furthermore, a case study is incorporated to provide new insights into the current AI methodologies in the domain. It also takes into account the important aspects of transparency and interpretability.

7.1 Conclusions

This section concludes the thesis with a summary of the objectives achieved.

In ***Chapter 3***, we proposed a multi-stage keyframe extraction framework and a novel polyp localization model, YcOLOn. Our framework reduces the burden of processing colonoscopy videos by eliminating redundant and uninformative frames. Besides reducing the processing time and effort, it enhances the polyp detection and localization performance. This framework is integrated with a GUI application that can help extract frames using a wide range of clickable options. In addition to the technical contributions, our work facilitates the research community with additional annotations of 49,136 polyp frames. These

annotations classify these polyp frames into informative/uninformative classes and are accessible on request.

In ***Chapter 4***, we proposed two solutions to deal with artifacts. One solution entirely translates an uninformative frame to an informative frame and the other solution performs segmentation of specular highlights. The translation process is performed using CycleGAN, a GAN-based architecture. Our approach successfully removed artifacts like ghost colors, fecal depositions and low illuminations. This translation is followed by a validation procedure which uses the YOLOv5 detector to localize polyps before and after translation. Experimental results show that more polyps were captured when translated frames were used. Hence, it could be inferred that our methodology enhanced the obscured polyp-specific features.

Similarly, in the second solution, a DWT-based encoder-decoder network is proposed to perform the segmentation of specular highlights. This particular artifact is handled using a separate framework as these highlights can even arise in visually clear and good-quality images. To extract clinical information from such images, we proposed a method to segment out specular highlights from them. Also, we modified one of the sub-bands in DWT to enhance our model's performance in overexposed regions. It is noteworthy that we used only the blue channel of the image as a specular component is easier to observe in this channel. Our approach performed superior to the state-of-the-art models even in overexposed regions.

In ***Chapter 5***, we provide a detailed summary and analysis of various techniques proposed by a total of 34 teams from two competitions. This study discussed two tasks (polyp segmentation, and algorithm efficiency) and three tasks (polyp segmentation, instrument segmentation, and transparency) of competitions from 2020 and 20221, respectively. These tasks are aimed at developing state-of-the-art algorithms including performance evaluation, resource utilization (efficiency) and transparency.

Further, considering the important aspects of the above competition platforms (such as fair comparison at the algorithm level in terms of standard metrics and interpretability), we focused on developing a multi-class dataset that can be utilized by the research community. Such a dataset would help conduct fair comparisons and reproducibility for different lesion types. Our dataset consists of 8000 GI images with 27 classes. These classes include normal, anatomical and pathological findings. We also provide experimental results using our dataset on baseline deep learning models.

In ***Chapter 6***, we proposed two approaches with different control mechanisms to

generate synthetic colonoscopy images. The first approach used an image-based input control map whereas the second method used a text-based control. The image-based input control map is obtained by overlapping the non-polyp images with a custom user-defined mask. Leveraging the non-polyp images which are relatively easy to access, we developed a diffusion-based method to convert non-polyp images into polyp images. Consequently, we obtained polyp frames along with ground truth (which were user-defined) without additional manual labeling effort. We further used the generated images to expand the training data for polyp segmentation. We observed that the segmentation performance improved by incorporating synthetic data with real data.

In the second approach, we proposed a diffusion-based method to generate a diverse set of polyp data that involves different pathology types (adenomatous/hyperplastic), and imaging techniques (NBI/WLI) and possesses good-quality. Our method used a text-controlled mechanism which takes a text prompt as input and outputs the desired image, matching the description provided in the text prompt. Further, we proposed a concept of cross-class label learning which allows the model to learn patterns from other classes. The generated images are combined with real images in different proportions and are used to perform polyp classification (adenomatous/hyperplastic). We observed that synthetic images obtained using a text-controlled mechanism along with cross-label learning enhanced the polyp classification performance.

7.2 Limitations of the Proposed Work

The proposed work achieves enhanced performance on the respective tasks and can contribute to different phases of the CAD system for clinical use. However, there are some limitations of the proposed work, which are discussed below.

Multiple stages and requirements of human involvement: The proposed work in Chapter 3 uses a multi-stage framework which creates opportunities for a single-stage system with similar or enhanced automation in dealing with uninformative frames. Our method requires some manual tuning based on the extent of low-quality or good-quality content in the input data. Although such manual management provides flexibility but at the same time increases overhead.

Persistence of some artifacts: The translation technique proposed work in Chapter 4 tackles various artifacts as desired, however, there are still some artifacts for which the method fails. These artifacts include motion blur and interlacing. This motivates the

research community to apply some blur removal algorithms or to develop a standalone system to deal with all artifacts using a single model.

Segmentation ground truth and sample count: The dataset proposed in Chapter 5 provides labels for multiple classes, however, it lacks annotations for segmentation. Also, there are a few classes for which the sample count is low as these anomalies are rare-to-find and difficult to capture during endoscopy.

Improper color retention in generated images: The synthetic data generation technique in Chapter 6 converts non-polyp images to polyp images and allows controlling the polyp location, size and shape. Additionally, the endoluminal scene structure of non-polyp frames remains intact. However, during the conversion process, the color is not retained in the generated images. Color could be an important attribute considered for domain shift problems.

7.3 Future Work Directions

While the dissertation has made significant progress in the endoscopy domain, several unresolved issues persist. Many of these are worth pursuing as future work and are discussed below.

Standalone system with fewer human interventions: For full automation and fewer human interventions for the process of keyframe selection, there is a need for a standalone system that considers all video attributes, and at the same time adapts to the input data. This adaptation would eliminate the need for any tuning of settings during inference. Also, the system can focus on improving generalizability capabilities for clinical use.

One system for several types of artifacts: As an endoscopy video is susceptible to many quality degrading factors, various artifacts may appear in the recordings. These artifacts have varied characteristics and handling them with a single model is difficult but an interesting problem. Such systems would be helpful in data pre-processing and indirectly could save manual effort.

Inpainting of specular highlights: The thesis provides a methodology to perform specularity segmentation but the process of inpainting to reconstruct the missing regions has not been discussed. This opens the possibility to develop an inpainting method.

Also, these two steps of specularity segmentation and inpainting can be combined in a single model or could be handled in a different manner, such as by using adversarial networks.

Unmapped or unlabeled data learning and color transfer: The existing works in synthetic image generation in endoscopy domain rely on a training dataset that contains input data mapped with its ground truth. However, obtaining such a fully mapped dataset or a labeled dataset could be very costly. Therefore, some methods could be proposed to generate synthetic data without such requirements. Also, the limitation of our work in Chapter 6 in retaining colors provides an opportunity to explore and develop some control mechanisms for colors in this domain.

Publications

Published/Accepted

[1] Vanshali Sharma, Pradipta Sasmal, MK Bhuyan, Pradip K Das, Yuji Iwahori, and Kunio Kasugai. "A multi-scale attention framework for automated polyp localization and keyframe extraction from colonoscopy videos". In: *IEEE Transactions on Automation Science and Engineering* (2023). URL: https://ieeexplore.ieee.org/document/10268934.

[2] Vanshali Sharma, Pradipta Sasmal, MK Bhuyan, and Pradip K Das. "Keyframe Selection from Colonoscopy Videos to Enhance Visualization for Polyp Detection". In: *2022 26th International Conference Information Visualisation (IV)*. IEEE. 2022, pp. 426–431. URL: https://ieeexplore.ieee.org/document/10017475.

[3] Vanshali Sharma, MK Bhuyan, and Pradip K Das. "Can adversarial networks make uninformative colonoscopy video frames clinically informative? (student abstract)". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 13. 2023, pp. 16322–16323. URL: https://doi.org/10.1609/aaai.v37i13.27021.

[4] Vanshali Sharma, MK Bhuyan, Pradip K Das, and Kangkana Bora. "A DWT-based encoder-decoder network for Specularity segmentation in colonoscopy images". In: *Multimedia Tools and Applications* (2023), pp. 1–20. URL: https://link.springer.com/article/10.1007/s11042-023-14564-1.

[5] Debesh Jha*, Vanshali Sharma*, Neethi Dasu, Nikhil Kumar Tomar, Steven Hicks, MK Bhuyan, Pradip K Das, Michael A Riegler, Pål Halvorsen, Ulas Bagci, et al. "GastroVision: A Multi-class Endoscopy Image Dataset for Computer Aided Gastrointestinal Disease Detection". In: *Workshop on Machine Learning for Multimodal Healthcare Data, co-located with ICML 2023*. Springer. 2023, pp. 125–140. URL: https://link.springer.com/chapter/10.1007/978-3-031-47679-2_10.

* Joint Contribution

[6] Vanshali Sharma, Abhishek Kumar, Debesh Jha, M.K. Bhuyan, Pradip K. Das, and Ulas Bagci. "ControlPolypNet: Towards Controlled Colon Polyp Synthesis for Improved Polyp Segmentation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. June 2024, pp. 2325–2334. URL: https://openaccess.thecvf.com/content/CVPR2024W/DCAMI/html/Sharma_ControlPolypNet_Towards_Controlled_Colon_Polyp_Synthesis_for_Improved_Polyp_Segmentation_CVPRW_2024_paper.html.

[7] Debesh Jha, Vanshali Sharma, Debapriya Banik, Debayan Bhattacharya, Kaushiki Roy, Steven A Hicks, Nikhil Kumar Tomar, Vajira Thambawita, Adrian Krenzer, Ge-Peng Ji, et al. "Validating polyp and instrument segmentation methods in colonoscopy through Medico 2020 and MedAI 2021 Challenges". In: *Medical Image Analysis* (2024). URL: https://arxiv.org/abs/2307.16262.

Under Review

[1] Vanshali Sharma, Debesh Jha, MK Bhuyan, Pradip K Das, and Ulas Bagci. *Generating Diverse Modality Colonoscopy Images Leveraging Cross-Class Labels*.

Out of Thesis

[1] Yashwanth Manjunatha, Vanshali Sharma, Yuji Iwahori, MK Bhuyan, Aili Wang, Akira Ouchi, and Yasuhiro Shimizu. "Lymph node detection in CT scans using modified U-Net with residual learning and 3D deep network". In: *International Journal of Computer Assisted Radiology and Surgery* 18.4 (2023), pp. 723–732. URL: https://link.springer.com/article/10.1007/s11548-022-02822-w.

[2] Promit Haldar, Vanshali Sharma, Yuji Iwahori, MK Bhuyan, Aili Wang, Haibin Wu, and Kunio Kasugai. "XGBoosted Binary CNNs for Multi-Class Classification of Colorectal Polyp Size". In: *IEEE Access* 11 (2023), pp. 128461–128472. URL: https://doi.org/10.1109/ACCESS.2023.3332826.

[3] Pradipta Sasmal, Vanshali Sharma, MK Bhuyan, Allam Jaya Prakash, Kiran Kumar Patro, Nagwan Abdel Samee, Hayam Alamro, Yuji Iwahori, Ryszard Tadeusiewicz, U Rajendra Acharya, et al. "Semi-supervised generative adversarial networks for improved colorectal polyp classification using histopathological images". In: *Information Sciences* (2023), p. 120033. URL: https://doi.org/10.1016/j.ins.2023.120033.

[4] Darshika Verma, Vanshali Sharma, and Pradip K Das. "Semi-supervised Polyp Classification in Colonoscopy Images Using GAN". In: *International Conference on Computer Vision and Image Processing*. Springer. 2023, pp. 39–51. URL: https://doi.org/10.1007/978-3-031-58535-7_4.

[5] Farhan Mohammad, Vanshali Sharma, and Pradip K Das. "Polyp Detection in Colonoscopy Images using Improved Deformable DETR". In: *TENCON 2022-2022 IEEE Region 10 Conference (TENCON)*. IEEE. 2022, pp. 1–6. URL: https://doi.org/10.1109/TENCON55691.2022.9977600.

[6] Nahush V Bhamre, Vanshali Sharma, Yuji Iwahori, MK Bhuyan, and Kunio Kasugai. "Colonoscopy Polyp Classification Adding Generated Narrow Band Imaging". In: *International Conference on Computer Vision and Image Processing*. Springer. 2022, pp. 322–334. URL: https://link.springer.com/chapter/10.1007/978-3-031-31417-9_25.

[7] Bhagath Parabattina, Phool Chandra, Vanshali Sharma, and Pradip K Das. "Voice-controlled assistance for robot navigation using android-based mobile devices". In: *2021 Sixth International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*. IEEE. 2021, pp. 21–25. URL: https://doi.org/10.1109/WiSPNET51692.2021.9419461.

Miscellaneous Details

GitHub Repositories and Dataset Link

- A Multi-scale Attention Framework for Polyp Localization and Keyframe Extraction from Colonoscopy Videos
- ControlPolypNet: Towards Controlled Colon Polyp Synthesis for Improved Polyp Segmentation
- GastroVision: A Multi-class Endoscopy Image Dataset for Computer Aided Gastrointestinal Disease Detection
- The *GastroVision* dataset can be downloaded using this link.

Services

- Reviewer: *MICCAI 2024, CVPR 2024, Scientifc Reports, Multimedia Systems (Springer), ICML Workshop on Machine Learning for Multimodal Healthcare Data 2023.*
- Volunteer: *AAAI Conference on Artificial Intelligence 2023, Washington DC, USA.*

Grants and Awards

- Awarded Travel Grants by *AAAI 2023, CVPR 2024, Research Week with Google 2023.*
- Awarded *INSPIRE Fellowship* by Department of Science and Technology (DST), Govt. of India.

Bibliography

[1] Kunio Doi. "Computer-aided diagnosis in medical imaging: historical review, current status and future potential". In: *Computerized medical imaging and graphics* 31.4-5 (2007), pp. 198–211.

[2] Shah Hussain, Iqra Mubeen, Niamat Ullah, Syed Shahab Ud Din Shah, Bakhtawar Abduljalil Khan, Muhammad Zahoor, Riaz Ullah, Farhat Ali Khan, and Mujeeb A Sultan. "Modern diagnostic imaging technique applications and risk factors in the medical field: A review". In: *BioMed Research International* 2022 (2022).

[3] Faridoddin Shariaty and Mojtaba Mousavi. "Application of CAD systems for the automatic detection of lung nodules". In: *Informatics in Medicine Unlocked* 15 (2019), p. 100173. ISSN: 2352-9148.

[4] Charlene H Chu, Simon Donato-Woodger, Shehroz S Khan, Rune Nyrup, Kathleen Leslie, Alexandra Lyn, Tianyu Shi, Andria Bianchi, Samira Abbasgholizadeh Rahimi, and Amanda Grenier. "Age-related bias and artificial intelligence: a scoping review". In: *Humanities and Social Sciences Communications* 10.1 (2023), pp. 1–17.

[5] Fadi Thabtah, Suhel Hammoud, Firuz Kamalov, and Amanda Gonsalves. "Data imbalance in classification: Experimental evaluation". In: *Information Sciences* 513 (2020), pp. 429–441. ISSN: 0020-0255.

[6] Melina Arnold, Christian C Abnet, Rachel E Neale, Jerome Vignat, Edward L Giovannucci, Katherine A McGlynn, and Freddie Bray. "Global burden of 5 major types of gastrointestinal cancer". In: *Gastroenterology* 159.1 (2020), pp. 335–349.

[7] Globocan. *Cancer Today*. 2020. URL: https://gco.iarc.fr/today/fact-sheets-cancers.

[8] Xiaoyong Yang, Qianxing Wei, Changhe Zhang, Kaibo Zhou, Li Kong, and Weiwei Jiang. "Colon polyp detection and segmentation based on improved MRCNN". In: *IEEE Transactions on Instrumentation and Measurement* 70 (2020), pp. 1–10.

[9] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. "Kvasir-SEG: A segmented polyp dataset". In: *Proceedings of the International Conference on Multimedia Modeling (MMM)*. 2020, pp. 451–462.

[10] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. "WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians". In: *Computerized medical imaging and graphics* 43 (2015), pp. 99–111.

[11] Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. "Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer". In: *International journal of computer assisted radiology and surgery* 9.2 (2014), pp. 283–293.

[12] Sharib Ali, Noha Ghatwary, Barbara Braden, Dominique Lamarque, Adam Bailey, Stefano Realdon, Renato Cannizzaro, Jens Rittscher, Christian Daul, and James East. "Endoscopy disease detection challenge 2020". In: *arXiv preprint arXiv:2003.03376* (2020).

[13] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Dag Johansen, Thomas De Lange, Pål Halvorsen, and Håvard D Johansen. "Resunet++: An advanced architecture for medical image segmentation". In: *In Proceedings of the IEEE International Symposium on Multimedia (ISM)*. 2019, pp. 225–2255.

[14] Debesh Jha, Michael A Riegler, Dag Johansen, Pål Halvorsen, and Håvard D Johansen. "DoubleUNet: A deep convolutional neural network for medical image segmentation". In: *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*. 2020, pp. 558–564.

[15] Chien-Hsiang Huang, Hung-Yu Wu, and Youn-Long Lin. "HarDNet-MSEG A Simple Encoder-Decoder Polyp Segmentation Neural Network that Achieves over 0.9 Mean Dice and 86 FPS". In: *arXiv preprint arXiv:2101.07172* (2021).

[16] Krushi Patel, Kaidong Li, Ke Tao, Quan Wang, Ajay Bansal, Amit Rastogi, and Guanghui Wang. "A comparative study on polyp classification using convolutional neural networks". In: *PloS one* 15.7 (2020), e0236452.

[17] Kaidong Li, Mohammad I Fathan, Krushi Patel, Tianxiao Zhang, Cuncong Zhong, Ajay Bansal, Amit Rastogi, Jean S Wang, and Guanghui Wang. "Colonoscopy Polyp

Detection and Classification: Dataset Creation and Comparative Evaluations". In: *arXiv preprint arXiv:2104.10824* (2021).

[18] Hayato Itoh, Masashi Misawa, Yuichi Mori, Masahiro Oda, Shin-Ei Kudo, and Kensaku Mori. *SUN Colonoscopy Video Database*. http://amed8k.sundatabase.org/. 2020.

[19] Jorge Bernal and Histace Aymeric. *MICCAI endoscopic vision challenge polyp detection and segmentation*. 2017.

[20] Vanshali Sharma, MK Bhuyan, and Pradip K Das. "Can adversarial networks make uninformative colonoscopy video frames clinically informative? (student abstract)". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 13. 2023, pp. 16322–16323. URL: https://doi.org/10.1609/aaai.v37i13.27021.

[21] Vanshali Sharma, MK Bhuyan, Pradip K Das, and Kangkana Bora. "A DWT-based encoder-decoder network for Specularity segmentation in colonoscopy images". In: *Multimedia Tools and Applications* (2023), pp. 1–20. URL: https://link.springer.com/article/10.1007/s11042-023-14564-1.

[22] Vanshali Sharma, Abhishek Kumar, Debesh Jha, M.K. Bhuyan, Pradip K. Das, and Ulas Bagci. "ControlPolypNet: Towards Controlled Colon Polyp Synthesis for Improved Polyp Segmentation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. June 2024, pp. 2325–2334. URL: https://openaccess.thecvf.com/content/CVPR2024W/DCAMI/html/Sharma_ControlPolypNet_Towards_Controlled_Colon_Polyp_Synthesis_for_Improved_Polyp_Segmentation_CVPRW_2024_paper.html.

[23] Vanshali Sharma, Debesh Jha, MK Bhuyan, Pradip K Das, and Ulas Bagci. *Generating Diverse Modality Colonoscopy Images Leveraging Cross-Class Labels*.

[24] Eileen Morgan, Melina Arnold, A Gini, V Lorenzoni, CJ Cabasag, Mathieu Laversanne, Jerome Vignat, Jacques Ferlay, Neil Murphy, and Freddie Bray. "Global burden of colorectal cancer in 2020 and 2040: incidence and mortality estimates from GLOBOCAN". In: *Gut* 72.2 (2023), pp. 338–344.

[25] R Lambert. "The Paris endoscopic classification of superficial neoplastic lesions: esophagus, stomach, and colon: November 30 to December 1, 2002". In: *Gastrointest Endosc* 58 (2003), S3–S43.

[26] Joshua C Obuch, Courtney M Pigott, and Dennis J Ahnen. “Sessile serrated polyps: detection, eradication, and prevention of the evil twin”. In: *Current treatment options in gastroenterology* 13 (2015), pp. 156–170.

[27] Kevin O Turner, Robert M Genta, and Amnon Sonnenberg. “Lesions of all types exist in colon polyps of all sizes”. In: *Official journal of the American College of Gastroenterology| ACG* 113.2 (2018), pp. 303–306.

[28] Samir Gupta, David Lieberman, Joseph C Anderson, Carol A Burke, Jason A Dominitz, Tonya Kaltenbach, Douglas J Robertson, Aasma Shaukat, Sapna Syngal, and Douglas K Rex. “Recommendations for follow-up after colonoscopy and polypectomy: a consensus update by the US Multi-Society Task Force on Colorectal Cancer”. In: *Official journal of the American College of Gastroenterology| ACG* 115.3 (2020), pp. 415–434.

[29] Sabina Beg, Ana Wilson, and Krish Ragunath. “The use of optical imaging techniques in the gastrointestinal tract”. In: *Frontline gastroenterology* 7.3 (2016), pp. 207–215.

[30] Mikihiro Fujiya, Kentaro Moriichi, Nobuhiro Ueno, Yusuke Saitoh, and Yutaka Kohgo. “Autofluorescence imaging for diagnosing intestinal disorders”. In: *Colonoscopy*. IntechOpen, 2011.

[31] Benjamin M Shandro, Khemraj Emrith, Gregory Slabaugh, Andrew Poullis, and Melvyn L Smith. “Optical imaging technology in colonoscopy: Is there a role for photometric stereo?” In: *World Journal of Gastrointestinal Endoscopy* 12.5 (2020), p. 138.

[32] Sun Young Park, Dustin Sargent, Inbar Spofford, Kirby G Vosburgh, A Yousif, et al. “A colon video analysis framework for polyp detection”. In: *IEEE Transactions on Biomedical Engineering* 59.5 (2012), pp. 1408–1418.

[33] Yixuan Yuan and Max Q-H Meng. “Hierarchical key frames extraction for WCE video”. In: *2013 IEEE International Conference on Mechatronics and Automation*. IEEE. 2013, pp. 225–229.

[34] JungHwan Oh, Sae Hwang, JeongKyu Lee, Wallapak Tavanapong, Johnny Wong, and Piet C de Groen. “Informative frame classification for endoscopy video”. In: *Medical Image Analysis* 11.2 (2007), pp. 110–127.

[35] Mirko Arnold, Anarta Ghosh, Gerard Lacey, Stephen Patchett, and Hugh Mulcahy. “Indistinct frame detection in colonoscopy videos”. In: *2009 13th International Machine Vision and Image Processing Conference*. IEEE. 2009, pp. 47–52.

[36] Baopu Li, Max Q-H Meng, and Qian Zhao. “Wireless capsule endoscopy video summary”. In: *2010 IEEE International Conference on Robotics and Biomimetics*. IEEE. 2010, pp. 454–459.

[37] Dimitrios K Iakovidis, Spyros Tsevas, and Andreas Polydorou. “Reduction of capsule endoscopy reading times by unsupervised image mining”. In: *Computerized Medical Imaging and Graphics* 34.6 (2010), pp. 471–478.

[38] Sandra Eliza Fontes De Avila, Ana Paula Brandao Lopes, Antonio da Luz Jr, and Arnaldo de Albuquerque Araújo. “VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method”. In: *Pattern recognition letters* 32.1 (2011), pp. 56–68.

[39] Naveed Ejaz, Irfan Mehmood, and Sung Wook Baik. “MRT letter: Visual attention driven framework for hysteroscopy video abstraction”. In: *Microscopy research and technique* 76.6 (2013), pp. 559–563.

[40] Engin Mendi, Coskun Bayrak, Songul Cecen, and Emre Ermisoglu. “Content-based management service for medical videos”. In: *Telemedicine and e-Health* 19.1 (2013), pp. 36–41.

[41] Mingyang Ma, Shaohui Mei, Shuai Wan, Zhiyong Wang, and Dagan Feng. “Video summarization via nonlinear sparse dictionary selection”. In: *IEEE Access* 7 (2019), pp. 11763–11774.

[42] Pradipta Sasmal, Avinash Paul, Manas Kamal Bhuyan, Yuji Iwahori, and Kunio Kasugai. “Extraction of Key-Frames From Endoscopic Videos by Using Depth Information”. In: *IEEE Access* 9 (2021), pp. 153004–153011.

[43] Mingyang Ma, Shaohui Mei, Shuai Wan, Zhiyong Wang, Zongyuan Ge, Vincent Lam, and Dagan Feng. “Keyframe extraction from laparoscopic videos via diverse and weighted dictionary selection”. In: *IEEE Journal of Biomedical and Health Informatics* 25.5 (2020), pp. 1686–1698.

[44] Constantinos Loukas, Christos Varytimidis, Konstantinos Rapantzikos, and Meletios A Kanakis. “Keyframe extraction from laparoscopic videos based on visual saliency

detection". In: *Computer Methods and Programs in Biomedicine* 165 (2018), pp. 13–23.

[45] Zhang Xu, Yu Tao, Zheng Wenfang, Lin Ne, Huang Zhengxing, Liu Jiquan, Hu Weiling, Duan Huilong, and Si Jianmin. "Upper gastrointestinal anatomy detection with multi-task convolutional neural networks". In: *Healthcare technology letters* 6.6 (2019), pp. 176–180.

[46] Ziyi Jin, Tianyuan Gan, Peng Wang, Zuoming Fu, Chongan Zhang, Qinglai Yan, Xueyong Zheng, Xiao Liang, and Xuesong Ye. "Deep learning for gastroscopic images: computer-aided techniques for clinicians". In: *BioMedical Engineering OnLine* 21.1 (2022), pp. 1–41.

[47] Shuai Wang, Yang Cong, Jun Cao, Yunsheng Yang, Yandong Tang, Huaici Zhao, and Haibin Yu. "Scalable gastroscopic video summarization via similar-inhibition dictionary selection". In: *Artificial intelligence in medicine* 66 (2016), pp. 1–13.

[48] Thomas Stehle, Roland Auer, Sebastian Gross, Alexander Behrens, Jonas Wulff, Til Aach, Ron Winograd, Christian Trautwein, and Jens Tischendorf. "Classification of colon polyps in NBI endoscopy using vascularization features". In: *Medical Imaging 2009: Computer-Aided Diagnosis*. Vol. 7260. International Society for Optics and Photonics. 2009, 72602S.

[49] Sebastian Gross, Stephan Palm, Jens JW Tischendorf, Alexander Behrens, Christian Trautwein, and Til Aach. "Automated classification of colon polyps in endoscopic image data". In: *Medical Imaging 2012: Computer-Aided Diagnosis*. Vol. 8315. International Society for Optics and Photonics. 2012, 83150W.

[50] Mojtaba Akbari, Majid Mohrekesh, Kayvan Najariani, Nader Karimi, Shadrokh Samavi, and SM Reza Soroushmehr. "Adaptive specular reflection detection and inpainting in colonoscopy video frames". In: *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE. 2018, pp. 3134–3138.

[51] Isabel N Figueiredo, Luis Pinto, Pedro N Figueiredo, and Richard Tsai. "Unsupervised segmentation of colonic polyps in narrow-band imaging data based on manifold representation of images and Wasserstein distance". In: *Biomedical Signal Processing and Control* 53 (2019), p. 101577.

[52] Baoxian Yu, Wanbing Chen, Qinghua Zhong, and Han Zhang. "Specular Highlight Detection Based on Color Distribution for Endoscopic Images". In: *Frontiers in Physics* 8 (2021), p. 575.

[53] Ranyang Li, Junjun Pan, Yaqing Si, Bin Yan, Yong Hu, and Hong Qin. "Specular reflections removal for endoscopic image sequences with adaptive-RPCA decomposition". In: *IEEE transactions on medical imaging* 39.2 (2019), pp. 328–340.

[54] Jorge Bernal, Javier Sánchez, and Fernando Vilarino. "Impact of image preprocessing methods on polyp localization in colonoscopy frames". In: *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2013, pp. 7350–7354.

[55] F Javier Sánchez, Jorge Bernal, Cristina Sánchez-Montes, Cristina Rodríguez de Miguel, and Gloria Fernández-Esparrach. "Bright spot regions segmentation and classification for specular highlights detection in colonoscopy videos". In: *Machine Vision and Applications* 28.8 (2017), pp. 917–936.

[56] Jorge Bernal, Javier Sánchez, and Fernando Vilarino. "Towards automatic polyp detection with a polyp appearance model". In: *Pattern Recognition* 45.9 (2012), pp. 3166–3182.

[57] Pradipta Sasmal, MK Bhuyan, Shashwata Gupta, and Yuji Iwahori. "Detection of Polyps in Colonoscopic Videos using Saliency Map based Modified Particle Filter". In: *IEEE Transactions on Instrumentation and Measurement* (2021).

[58] Vajira Thambawita, Debesh Jha, Hugo Lewi Hammer, Håvard D Johansen, Dag Johansen, Pål Halvorsen, and Michael A Riegler. "An extensive study on cross-dataset bias and evaluation metrics interpretation for machine learning applied to gastrointestinal tract abnormality classification". In: *ACM Transactions on Computing for Healthcare* 1.3 (2020), pp. 1–29.

[59] Ishak Pacal, Ahmet Karaman, Dervis Karaboga, Bahriye Akay, Alper Basturk, Ufuk Nalbantoglu, and Seymanur Coskun. "An efficient real-time colonic polyp detection with YOLO algorithms trained by using negative samples and large datasets". In: *Computers in biology and medicine* 141 (2022), p. 105031.

[60] Zhiqin Qian, Yi Lv, Dongyuan Lv, Huijun Gu, Kunyu Wang, Wenjun Zhang, and Madan M Gupta. "A New Approach to Polyp Detection by Pre-Processing of Images and Enhanced Faster R-CNN". In: *IEEE Sensors Journal* 21.10 (2020), pp. 11374–11381.

[61] Yu Tian, Leonardo Zorron Cheng Tao Pu, Yuyuan Liu, Gabriel Maicas, Johan W Verjans, Alastair D Burt, Seon Ho Shin, Rajvinder Singh, and Gustavo Carneiro.

"Detecting, localising and classifying polyps from colonoscopy videos using deep learning". In: *arXiv preprint arXiv:2101.03285* (2021).

[62] Debesh Jha, Sharib Ali, Nikhil Kumar Tomar, Håvard D Johansen, Dag Johansen, Jens Rittscher, Michael A Riegler, and Pål Halvorsen. "Real-time polyp detection, localization and segmentation in colonoscopy using deep learning". In: *IEEE Access* 9 (2021), pp. 40496–40510.

[63] Jorge Bernal, Joan Manel Núñez, F Javier Sánchez, and Fernando Vilariño. "Polyp segmentation method in colonoscopy videos by means of MSA-DOVA energy maps calculation". In: *Proceedings of the Workshop on Clinical Image-Based Procedures*. 2014, pp. 41–49.

[64] Mojtaba Akbari, Majid Mohrekesh, Ebrahim Nasr-Esfahani, SM Reza Soroushmehr, Nader Karimi, Shadrokh Samavi, and Kayvan Najarian. "Polyp segmentation in colonoscopy images using fully convolutional network". In: *Proceedings of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2018, pp. 69–72.

[65] Patrick Brandao, Odysseas Zisimopoulos, Evangelos Mazomenos, Gastone Ciuti, Jorge Bernal, Marco Visentini-Scarzanella, Arianna Menciassi, Paolo Dario, Anastasios Koulaouzidis, Alberto Arezzo, et al. "Towards a computed-aided diagnosis system in colonoscopy: automatic polyp segmentation using convolution neural networks". In: *Journal of Medical Robotics Research* 3.02 (2018), p. 1840002.

[66] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer. 2015, pp. 234–241.

[67] Debesh Jha, Pia H Smedsrud, Dag Johansen, Thomas de Lange, Håvard D Johansen, Pål Halvorsen, and Michael A Riegler. "A comprehensive study on Colorectal Polyp Segmentation With ResUNet++, Conditional Random Field and Test-Time Augmentation". In: *IEEE journal of Biomedical and Health Informatics* 25.6 (2021), pp. 2029–2040.

[68] Debesh Jha, Sharib Ali, Håvard D Johansen, Dag D Johansen, Jens Rittscher, Michael A Riegler, and Pål Halvorsen. "Real-Time Polyp Detection, Localisation and Segmentation in Colonoscopy Using Deep Learning". In: *arXiv preprint arXiv:2011.07631* (2020).

[69] Michael Yeung, Evis Sala, Carola-Bibiane Schönlieb, and Leonardo Rundo. "Focus U-Net: A novel dual attention-gated CNN for polyp segmentation during colonoscopy". In: *Computers in biology and medicine* 137 (2021), p. 104815.

[70] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. "Pranet: Parallel reverse attention network for polyp segmentation". In: *Proceedings of the International conference on medical image computing and computer-assisted intervention (MICCAI)*. 2020, pp. 263–273.

[71] Guanghui Yue, Wanwan Han, Bin Jiang, Tianwei Zhou, Runmin Cong, and Tianfu Wang. "Boundary Constraint Network with Cross Layer Feature Integration for Polyp Segmentation". In: *IEEE Journal of Biomedical and Health Informatics* (2022).

[72] Na Ta, Haipeng Chen, Yingda Lyu, and Taosuo Wu. "BLE-Net: boundary learning and enhancement network for polyp segmentation". In: *Multimedia Systems* (2022), pp. 1–14.

[73] Bo Dong, Wenhai Wang, Deng-Ping Fan, Jinpeng Li, Huazhu Fu, and Ling Shao. "Polyp-PVT: Polyp Segmentation with Pyramid Vision Transformers". In: *arXiv preprint arXiv:2108.06932* (2021).

[74] Nguyen Thanh Duc, Nguyen Thi Oanh, Nguyen Thi Thuy, Tran Minh Triet, and Viet Sang Dinh. "Colonformer: An efficient transformer based method for colon polyp segmentation". In: *IEEE Access* 10 (2022), pp. 80575–80586.

[75] Wenchao Zhang, Chong Fu, Yu Zheng, Fangyuan Zhang, Yanli Zhao, and Chiu-Wing Sham. "HSNet: A hybrid semantic network for polyp segmentation". In: *Computers in Biology and Medicine* 150 (2022), p. 106173.

[76] Abhishek Srivastava, Sukalpa Chanda, Debesh Jha, Umapada Pal, and Sharib Ali. "GMSRF-Net: An improved generalizability with global multi-scale residual fusion network for polyp segmentation". In: *Proceedings of the 2022 26th International Conference on Pattern Recognition (ICPR)*. 2022, pp. 4321–4327.

[77] Ruiwei Feng, Biwen Lei, Wenzhe Wang, Tingting Chen, Jintai Chen, Danny Z Chen, and Jian Wu. "SSN: A stair-shape network for real-time polyp segmentation in colonoscopy images". In: *Proceedings of the 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. 2020, pp. 225–229.

[78] Sirojbek Safarov and Taeg Keun Whangbo. “A-DenseUNet: Adaptive densely connected UNet for polyp segmentation in colonoscopy images with atrous convolution”. In: *Sensors* 21.4 (2021), p. 1441.

[79] Nikhil Kumar Tomar, Debesh Jha, Ulas Bagci, and Sharib Ali. “TGANet: Text-guided attention for improved polyp segmentation”. In: *arXiv preprint arXiv:2205.04280* (2022).

[80] Huisi Wu, Zebin Zhao, Jiafu Zhong, Wei Wang, Zhenkun Wen, and Jing Qin. “PolypSeg+: A Lightweight Context-Aware Network for Real-Time Polyp Segmentation”. In: *IEEE Transactions on Cybernetics* (2022).

[81] Georg Wimmer, Michael Gadermayr, Roland Kwitt, Michael Häfner, Dorit Merhof, and Andreas Uhl. “Evaluation of i-scan virtual chromoendoscopy and traditional chromoendoscopy for the automated diagnosis of colonic polyps”. In: *International Workshop on Computer-Assisted and Robotic Endoscopy*. Springer. 2016, pp. 59–71.

[82] Andreas Uhl, Georg Wimmer, and Michael Hafner. “Shape and size adapted local fractal dimension for the classification of polyps in HD colonoscopy”. In: *2014 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2014, pp. 2299–2303.

[83] Michael Häfner, Toru Tamaki, Shinji Tanaka, Andreas Uhl, Georg Wimmer, and Shigeto Yoshida. “Local fractal dimension based approaches for colonic polyp classification”. In: *Medical image analysis* 26.1 (2015), pp. 92–107.

[84] Georg Wimmer, Andreas Uhl, and Michael Häfner. “A novel filterbank especially designed for the classification of colonic polyps”. In: *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE. 2016, pp. 2150–2155.

[85] Eduardo Ribeiro, Andreas Uhl, Georg Wimmer, and Michael Häfner. “Exploring deep learning and transfer learning for colonic polyp classification”. In: *Computational and mathematical methods in medicine* 2016 (2016).

[86] Kangkana Bora, MK Bhuyan, Kunio Kasugai, Saurav Mallik, and Zhongming Zhao. “Computational learning of features for automated colonic polyp classification”. In: *Scientific Reports* 11.1 (2021), pp. 1–16.

[87] Nahush V Bhamre, Vanshali Sharma, Yuji Iwahori, MK Bhuyan, and Kunio Kasugai. “Colonoscopy Polyp Classification Adding Generated Narrow Band Imaging”. In: *International Conference on Computer Vision and Image Processing*. Springer. 2022, pp. 322–334.

[88] Hanna Borgli, Vajira Thambawita, Pia H Smedsrud, Steven Hicks, Debesh Jha, Sigrun L Eskeland, Kristin Ranheim Randel, Konstantin Pogorelov, Mathias Lux, Duc Tien Dang Nguyen, et al. "HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy". In: *Scientific Data* 7.1 (2020), pp. 1–14.

[89] Pia H Smedsrud, Vajira Thambawita, Steven A Hicks, Henrik Gjestang, Oda Olsen Nedrejord, Espen Næss, Hanna Borgli, Debesh Jha, Tor Jan Derek Berstad, Sigrun L Eskeland, et al. "Kvasir-Capsule, a video capsule endoscopy dataset". In: *Scientific Data* 8.1 (2021), pp. 1–10.

[90] Konstantin Pogorelov et al. "Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection". In: *In Proceedings of the ACM Multim. Sys. Conf. (MMSys)*. 2017, pp. 164–169.

[91] Sharib Ali et al. "Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy". In: *Medical Image Analysis* 70 (2021), p. 102002.

[92] Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. "Automated polyp detection in colonoscopy videos using shape and context information". In: *IEEE Trans. Med. Imag* 35.2 (2015), pp. 630–644.

[93] Anastasios Koulaouzidis, Dimitris K Iakovidis, Diana E Yung, Emanuele Rondonotti, Uri Kopylov, John N Plevris, Ervin Toth, Abraham Eliakim, Gabrielle Wurm Johansson, Wojciech Marlicz, et al. "KID Project: an internet-based digital video atlas of capsule endoscopy for research purposes". In: *Endoscopy international open* 5.6 (2017), E477.

[94] Sharib Ali, Debesh Jha, Noha Ghatwary, Stefano Realdon, Renato Cannizzaro, Osama E Salem, Dominique Lamarque, Christian Daul, Michael A Riegler, Kim V Anonsen, et al. "A multi-centre polyp detection and segmentation dataset for generalisability assessment". In: *Scientific Data* 10.1 (2023), p. 75.

[95] Masashi Misawa, Shin-ei Kudo, Yuichi Mori, Kinichi Hotta, Kazuo Ohtsuka, Takahisa Matsuda, Shoichi Saito, Toyoki Kudo, Toshiyuki Baba, Fumio Ishida, et al. "Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video)". In: *Gastrointestinal endoscopy* 93.4 (2021), pp. 960–967.

[96] Pablo Mesejo, Daniel Pizarro, Armand Abergel, Olivier Rouquette, Sylvain Beorchia, Laurent Poincloux, and Adrien Bartoli. "Computer-aided classification of gastrointestinal lesions in regular colonoscopy". In: *IEEE transactions on medical imaging* 35.9 (2016), pp. 2051–2063.

[97] Younghak Shin, Hemin Ali Qadir, and Ilangko Balasingham. "Abnormal colon polyp image synthesis using conditional adversarial networks for improved detection performance". In: *IEEE Access* 6 (2018), pp. 56007–56017.

[98] Hemin Ali Qadir, Ilangko Balasingham, and Younghak Shin. "Simple U-net based synthetic polyp image generation: Polyp to negative and negative to polyp". In: *Biomedical Signal Processing and Control* 74 (2022), p. 103491.

[99] Jan Andre Fagereng, Vajira Thambawita, Andrea M Storås, Sravanthi Parasa, Thomas De Lange, Pål Halvorsen, and Michael A Riegler. "PolypConnect: Image inpainting for generating realistic gastrointestinal tract images with polyps". In: *2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE. 2022, pp. 66–71.

[100] Pradipta Sasmal, MK Bhuyan, Sourav Sonowal, Yuji Iwahori, and Kunio Kasugai. "Improved endoscopic polyp classification using GAN generated synthetic data augmentation". In: *2020 IEEE Applied Signal Processing Conference (ASPCON)*. IEEE. 2020, pp. 247–251.

[101] Prince Ebenezer Adjei, Zenebe Markos Lonseko, Wenju Du, Han Zhang, and Nini Rao. "Examining the effect of synthetic data augmentation in polyp detection and segmentation". In: *International Journal of Computer Assisted Radiology and Surgery* 17.7 (2022), pp. 1289–1302.

[102] Fan He, Sizhe Chen, Shuaiyi Li, Lu Zhou, Haiqin Zhang, Haixia Peng, and Xiaolin Huang. "Colonoscopic image synthesis for polyp detector enhancement via gan and adversarial training". In: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2021, pp. 1887–1891.

[103] Ataher Sams and Homaira Huda Shomee. "GAN-Based Realistic Gastrointestinal Polyp Image Synthesis". In: *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2022, pp. 1–4.

[104] Mayank Golhar, Taylor L Bobrow, Saowanee Ngamruengphong, and Nicholas J Durr. "GAN Inversion for Data Augmentation to Improve Colonoscopy Lesion Classification". In: *arXiv preprint arXiv:2205.02840* (2022).

[105] Roman Macháček, Leila Mozaffari, Zahra Sepasdar, Sravanthi Parasa, Pål Halvorsen, Michael A Riegler, and Vajira Thambawita. "Mask-conditioned latent diffusion for generating gastrointestinal polyp images". In: *arXiv preprint arXiv:2304.05233* (2023).

[106] Alexander K Pishva, Vajira Thambawita, Jim Torresen, and Steven A Hicks. "Re-Polyp: A Framework for Generating Realistic Colon Polyps with Corresponding Segmentation Masks using Diffusion Models". In: *2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE. 2023, pp. 47–52.

[107] Yuhao Du, Yuncheng Jiang, Shuangyi Tan, Xusheng Wu, Qi Dou, Zhen Li, Guanbin Li, and Xiang Wan. "ArSDM: Colonoscopy Images Synthesis with Adaptive Refinement Semantic Diffusion Models". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 339–349.

[108] Laura C Seeff, Thomas B Richards, Jean A Shapiro, Marion R Nadel, Diane L Manninen, Leslie S Given, Fred B Dong, Linda D Winges, and Matthew T McKenna. "How many endoscopies are performed for colorectal cancer screening? Results from CDC's survey of endoscopic capacity". In: *Gastroenterology* 127.6 (2004), pp. 1670–1677.

[109] Yixuan Yuan, Wenjian Qin, Bulat Ibragimov, Guanglei Zhang, Bin Han, Max Q-H Meng, and Lei Xing. "Densely connected neural network with unbalanced discriminant and category sensitive constraints for polyp recognition". In: *IEEE Transactions on Automation Science and Engineering* 17.2 (2019), pp. 574–583.

[110] Farhan Mohammad, Vanshali Sharma, and Pradip K Das. "Polyp Detection in Colonoscopy Images using Improved Deformable DETR". In: *TENCON 2022-2022 IEEE Region 10 Conference (TENCON)*. IEEE. 2022, pp. 1–6.

[111] Vanshali Sharma, MK Bhuyan, and Pradip K Das. "Can Adversarial Networks Make Uninformative Colonoscopy Video Frames Clinically Informative?" In: *arXiv preprint arXiv:2304.02152* (2023).

[112] Xiao Jia, Xiaochun Mai, Yi Cui, Yixuan Yuan, Xiaohan Xing, Hyunseok Seo, Lei Xing, and Max Q-H Meng. "Automatic polyp recognition in colonoscopy images using deep learning and two-stage pyramidal feature prediction". In: *IEEE Transactions on Automation Science and Engineering* 17.3 (2020), pp. 1570–1584.

[113] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. "Densely connected convolutional networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708.

[114] Renting Liu, Zhaorong Li, and Jiaya Jia. "Image partial blur detection and classification". In: *2008 IEEE conference on computer vision and pattern recognition*. IEEE. 2008, pp. 1–8.

[115] Said Pertuz, Domenec Puig, and Miguel Angel Garcia. "Analysis of focus measure operators for shape-from-focus". In: *Pattern Recognition* 46.5 (2013), pp. 1415–1432.

[116] Shui-Hua Wang and Yu-Dong Zhang. "DenseNet-201-based deep neural network with composite learning factor and precomputation for multiple sclerosis classification". In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 16.2s (2020), pp. 1–19.

[117] Xiang Yu, Nianyin Zeng, Shuai Liu, and Yu-Dong Zhang. "Utilization of DenseNet201 for diagnosis of breast abnormality". In: *Machine Vision and Applications* 30.7 (2019), pp. 1135–1144.

[118] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[119] *YOLOv5*. https://github.com/ultralytics/yolov5.git/. Accessed: 25 Feb 2022.

[120] Yimian Dai, Fabian Gieseke, Stefan Oehmcke, Yiquan Wu, and Kobus Barnard. "Attentional feature fusion". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2021, pp. 3560–3569.

[121] Hayato Itoh, Masashi Misawa, Yuichi Mori, Masahiro Oda, Shin-Ei Kudo, and Kensaku Mori. *SUN Colonoscopy Video Database*. http://amed8k.sundatabase.org/. 2020.

[122] Quentin Angermann, Jorge Bernal, Cristina Sánchez-Montes, Maroua Hammami, Gloria Fernández-Esparrach, Xavier Dray, Olivier Romain, F Javier Sánchez, and Aymeric Histace. "Towards real-time polyp detection in colonoscopy videos: Adapting still frame-based methodologies for video sequences analysis". In: *Computer Assisted and Robotic Endoscopy and Clinical Image-Based Procedures*. Springer, 2017, pp. 29–41.

[123] Jorge Bernal, Aymeric Histace, Marc Masana, Quentin Angermann, Cristina Sánchez-Montes, Cristina Rodriguez, Maroua Hammami, Ana Garcia-Rodriguez, Henry Córdova, Olivier Romain, et al. "Polyp detection benchmark in colonoscopy videos using gtcreator: A novel fully configurable tool for easy and fast annotation of image databases". In: *Proceedings of 32nd CARS conference*. 2018.

[124] Yasin Almalioglu, Kutsev Bengisu Ozyoruk, Abdulkadir Gokce, Kagan Incetan, Guliz Irem Gokceler, Muhammed Ali Simsek, Kivanc Ararat, Richard J Chen, Nicholas J Durr, Faisal Mahmood, et al. "EndoL2H: deep super-resolution for capsule endoscopy". In: *IEEE Transactions on Medical Imaging* 39.12 (2020), pp. 4297–4309.

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

[126] Ultralytics. *YOLOv5*. https://github.com/ultralytics/yolov5.git/. Accessed: 25 Feb 2022. 2020.

[127] Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng-Ann Heng. "H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes". In: *IEEE transactions on medical imaging* 37.12 (2018), pp. 2663–2674.

[128] Zitao Zeng, Weihao Xie, Yunzhe Zhang, and Yao Lu. "RIC-Unet: An improved neural network based on Unet for nuclei segmentation in histology images". In: *Ieee Access* 7 (2019), pp. 21420–21428.

[129] MR Pavan Kumar and Prabhu Jayagopal. "Generative adversarial networks: a survey on applications and challenges". In: *International Journal of Multimedia Information Retrieval* 10.1 (2021), pp. 1–24.

[130] Ekam Singh Chahal, Aarya Patel, Ayush Gupta, Archana Purwar, et al. "Unet based Xception Model for Prostate Cancer Segmentation from MRI Images". In: *Multimedia Tools and Applications* (2021), pp. 1–17.

[131] Ryuhei Hamaguchi, Aito Fujita, Keisuke Nemoto, Tomoyuki Imaizumi, and Shuhei Hikosaka. "Effective use of dilated convolutions for segmenting small object instances in remote sensing imagery". In: *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE. 2018, pp. 1442–1450.

[132] Pengju Liu, Hongzhi Zhang, Kai Zhang, Liang Lin, and Wangmeng Zuo. "Multi-level wavelet-CNN for image restoration". In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2018, pp. 773–782.

[133] Fisher Yu and Vladlen Koltun. "Multi-scale context aggregation by dilated convolutions". In: *arXiv preprint arXiv:1511.07122* (2015).

[134] Peiyun Hu and Deva Ramanan. "Finding tiny faces". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 951–959.

[135] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. "Object detectors emerge in deep scene cnns". In: *arXiv preprint arXiv:1412.6856* (2014).

[136] Krystian Pyka. "Wavelet-based local contrast enhancement for satellite, aerial and close range images". In: *Remote Sensing* 9.1 (2017), p. 25.

[137] R Priyadharsini, T Sree Sharmila, and V Rajendran. "Underwater acoustic image enhancement using wavelet and KL transform". In: *2015 International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*. IEEE. 2015, pp. 563–567.

[138] VB Prasath. "Polyp detection and segmentation from video capsule endoscopy: A review". In: *Journal of Imaging* 3.1 (2017), p. 1.

[139] Pradipta Sasmal, Manas Kamal Bhuyan, Kangkana Bora, Yuji Iwahori, and Kunio Kasugai. "Colonoscopic Image Polyp Classification Using Texture Features". In: *International Conference on Pattern Recognition and Machine Intelligence*. Springer. 2019, pp. 96–101.

[140] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. "Encoder-decoder with atrous separable convolution for semantic image segmentation". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 801–818.

[141] Ahana Roy Choudhury, Rami Vanguri, Sachin R Jambawalikar, and Piyush Kumar. "Segmentation of brain tumors using DeepLabv3+". In: *International MICCAI Brainlesion Workshop*. Springer. 2018, pp. 154–167.

[142] Muyi Sun, Guanhong Zhang, Hao Dang, Xingqun Qi, Xiaoguang Zhou, and Qing Chang. "Accurate gastric cancer segmentation in digital pathology images using deformable convolution and multi-scale embedding networks". In: *IEEE access* 7 (2019), pp. 75530–75541.

[143] Jorge Bernal and Histace Aymeric. *Gastrointestinal Image ANAlysis (GIANA) Angiodysplasia D&L challenge*. https://endovissub2017-giana.grand-challenge.org/home/. Accessed: 2017-11-20. 2017.

[144] Sharib Ali, Noha Ghatwary, Debesh Jha, Ece Isik-Polat, Gorkem Polat, Chen Yang, Wuyang Li, Adrian Galdran, Miguel-Ángel González Ballester, Vajira Thambawita, et al. "Assessing generalisability of deep learning-based polyp detection and segmentation methods through a computer vision challenge". In: *arXiv preprint arXiv:2202.12031* (2022).

[145] Steven Hicks, Debesh Jha, Vajira Thambawita, Hugo Hammer, Thomas de Lange, Sravanthi Parasa, Michael Riegler, and Pål Halvorsen. "Medico multimedia task at mediaeval 2021: Transparency in medical image segmentation". In: *Proceedings of MediaEval 2021 CEUR Workshop*. 2021, pp. 1–2.

[146] Debesh Jha, Sharib Ali, Krister Emanuelsen, Steven A Hicks, Vajira Thambawita, Enrique Garcia-Ceja, Michael A Riegler, Thomas de Lange, Peter T Schmidt, Håvard D Johansen, et al. "Kvasir-instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy". In: *MultiMedia Modeling: 27th International Conference, MMM 2021, Prague, Czech Republic, June 22–24, 2021, Proceedings, Part II 27*. 2021, pp. 218–229.

[147] Saurab Rauniyar, Vabesh Kumar Jha, Ritika Kumari Jha, Debesh Jha, and Ashish Rauniyar. "Improving Polyp Segmentation in Colonoscopy using Deep Learning". In: *Nordic Machine Intelligence* 1.1 (Nov. 2021), pp. 35–37.

[148] Debapriya Banik, Kaushiki Roy, and Debotosh Bhattacharjee. "EM-Net: An Efficient M-Net for segmentation of surgical instruments in colonoscopy frames". In: *Nordic Machine Intelligence* 1.1 (2021), pp. 14–16.

[149] Syed Muhammad Faraz Ali, Muhammad Taha Khan, Syed Unaiz Haider, Talha Ahmed, Zeshan Khan, and Muhammad Atif Tahir. "Depth-Wise Separable Atrous Convolution for Polyps Segmentation in Gastro-Intestinal Tract." In: *In Proceedings of the MediaEval*. 2020, pp. 1–3.

[150] Sabari Nathan and Suganya Ramamoorthy. “Efficient Supervision Net: Polyp Segmentation Using EfficientNet and Attention Unit.” In: *In Proceedings of the MediaEval*. 2020, pp. 1–3.

[151] Saruar Alam, Nikhil Kumar Tomar, Aarati Thakur, Debesh Jha, and Ashish Rauniyar. “Automatic polyp segmentation using u-net-resnet50”. In: *arXiv preprint arXiv:2012.15247* (2020).

[152] Awadelrahman Ahmed and MA Ali. “Generative Adversarial Networks for Automatic Polyp Segmentation”. In: *arXiv preprint arXiv:2012.06771* (2020).

[153] Quoc-Huy Trinh, Minh-Van Nguyen, Thiet-Gia Huynh, and Minh-Triet Tran. “HCMUS-Juniors 2020 at Medico Task in MediaEval 2020: Refined Deep Neural Network and U-Net for Polyps Segmentation.” In: *In Proceedings of the MediaEval*. 2020, pp. 1–3.

[154] Adrian Krenzer and Frank Puppe. “Bigger Networks are not Always Better: Deep Convolutional Neural Networks for Automated Polyp Segmentation.” In: *In Proceedings of the MediaEval*. 2020, pp. 1–3.

[155] Shruti Shrestha, Bishesh Khanal, and Sharib Ali. “Ensemble U-Net Model for Efficient Polyp Segmentation.” In: *In Proceedings of the MediaEval*. 2020, pp. 1–3.

[156] Jaeyong Kang and Jeonghwan Gwak. “KD-ResUNet++: Automatic Polyp Segmentation via Self-Knowledge Distillation.” In: *In Proceedings of the MediaEval*. 2020, pp. 1–3.

[157] Rabindra Khadka. “Transfer of Knowledge: Fine-tuning for Polyp Segmentation with Attention.” In: *In Proceedings of the MediaEval*. 2020, pp. 1–3.

[158] Debapriya Banik and Debotosh Bhattacharjee. “Deep Conditional Adversarial Learning for Polyp Segmentation.” In: *In Proceedings of the MediaEval*. 2020, pp. 1–3.

[159] Ge-Peng Ji, Deng-Ping Fan, Tao Zhou, Geng Chen, Huazhu Fu, and Ling Shao. “Automatic Polyp Segmentation via Parallel Reverse Attention Network.” In: *In Proceedings of the MediaEval*. 2020, pp. 1–3.

[160] Tien-Phat Nguyen, Tan-Cong Nguyen, Gia-Han Diep, Minh-Quan Le, Hoang-Phuc Nguyen-Dinh, Hai-Dang Nguyen, and Minh-Triet Tran. “HCMUS at Medico Automatic Polyp Segmentation Task 2020: PraNet and ResUnet++ for Polyps Segmentation.” In: *In Proceedings of the MediaEval*. 2020, pp. 1–3.

[161] George Batchkala and Sharib Ali. “Real-Time Polyp Segmentation Using U-Net with IoU Loss.” In: *In Proceedings of the MediaEval*. 2020, pp. 1–3.

[162] Sahadev Poudel and Sang-Woong Lee. “Automatic Polyp Segmentation Using Channel-Spatial Attention with Deep Supervision.” In: *In Proceedings of the MediaEval*. 2020, pp. 1–3.

[163] Vajira Thambawita, Steven Hicks, Pål Halvorsen, and Michael A Riegler. “Pyramid-Focus-Augmentation: Medical Image Segmentation with Step-Wise Focus”. In: *arXiv preprint arXiv:2012.07430* (2020).

[164] Cai-Wu Maxwell Hwang, Kao-Shing Hwang, Yong Si Xu, and Chien-Hsing Wu. “A Temporal-Spatial Attention Model for Medical Image Detection”. In: *In Proceedings of the MediaEval*. 2020, pp. 1–3.

[165] Nikhil Kumar Tomar. “Automatic Polyp Segmentation using Fully Convolutional Neural Network”. In: *arXiv preprint arXiv:2101.04001* (2021).

[166] Ashar Mirza and Rishav Kumar Rajak. “Segmentation of Polyp Instruments using UNet based deep learning model”. In: *Nordic Machine Intelligence* 1.1 (Nov. 2021), pp. 44–46.

[167] Ayush Somani, Divij Singh, Dilip Prasad, and Alexander Horsch. “T-MIS: Transparency Adaptation in Medical Image Segmentation”. In: *Nordic Machine Intelligence* 1.1 (2021), pp. 11–13.

[168] Debayan Bhattacharya, Christian Betz, Dennis Eggert, and Alexander Schlaefer. “Dual parallel reverse attention edge network : DPRA-EdgeNet”. In: *Nordic Machine Intelligence* 1.1 (2021), pp. 8–10.

[169] Mahmood Haithami, Amr Ahmed, Iman Yi Liao, and Hamid Jalab. “Employing GRU to combine feature maps in DeeplabV3 for a better segmentation model”. In: *Nordic Machine Intelligence* 1.1 (Nov. 2021), pp. 29–31.

[170] Bo Dong, Wenhai Wang, and Jinpeng Li. “Transformer Based Multi-model Fusion for Medical Image Segmentation”. In: *Nordic Machine Intelligence* 1.1 (Nov. 2021), pp. 50–52.

[171] Yung-Han Chen, Pei-Hsuan Kuo, Yi-Zeng Fang, and Wei-Lin Wang. “More Birds in the Hand -Medical Image Segmentation using a Multi-Model Ensemble Framework”. In: *Nordic Machine Intelligence* 1.1 (Nov. 2021), pp. 23–25.

[172] Sahadev Poudel and Sang-Woong Lee. “Explainable U-Net model forMedical Image Segmentation”. In: *Nordic Machine Intelligence* 1.1 (Nov. 2021), pp. 41–43.

[173] Awadelrahman Ahmed and Leen AM Ali. “Explainable Medical Image Segmentation via Generative Adversarial Networks and Layer-wise Relevance Propagation”. In: *arXiv preprint arXiv:2111.01665* (2021).

[174] YuCheng Chou. “Automatic Polyp and Instrument Segmentation in MedAI-2021”. In: *Nordic Machine Intelligence* 1.1 (Nov. 2021), pp. 17–19.

[175] Nefeli Panagiota Tzavara and Bjørn-Jostein Singstad. “Transfer Learning in Polyp and Endoscopic Tool Segmentation from Colonoscopy Images”. In: *Nordic Machine Intelligence* 1.1 (Nov. 2021), pp. 32–34.

[176] Adrian Galdran. “Polyp and Surgical Instrument Segmentation with Double Encoder-Decoder Networks”. In: *Nordic Machine Intelligence* 1.1 (2021), pp. 5–7.

[177] Arvind Keprate and Sumit Pandey. “Kvasir-Instruments and Polyp Segmentation Using UNet”. In: *Nordic Machine Intelligence* 1.1 (Nov. 2021), pp. 26–28.

[178] Michael Yeung. “Attention U-Net ensemble for interpretable polyp and instrument segmentation”. In: *Nordic Machine Intelligence* 1.1 (Nov. 2021), pp. 47–49.

[179] Sharib Ali and Nikhil K Tomar. “Iterative deep learning for improved segmentation of endoscopic images”. In: *Nordic Machine Intelligence* 1.1 (2021), pp. 38–40.

[180] S Ali et al. “An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy”. In: *Sci. Rep* (2020), pp. 1–21.

[181] Ping Chao, Chao-Yang Kao, Yu-Shan Ruan, Chien-Hsiang Huang, and Youn-Long Lin. “Hardnet: A low memory traffic network”. In: *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*. 2019, pp. 3552–3561.

[182] Yijun Yang, Zhaohu Xing, and Lei Zhu. “Vivim: a video vision mamba for medical video object segmentation”. In: *arXiv preprint arXiv:2401.14168* (2024).

[183] Pellegrino Crafa and Salvador J Diaz-Cano. “Changes in Colonic Structure and Mucosal Inflammation”. In: *Colonic Diverticular Disease*. 2022, pp. 41–61.

[184] Mingxing Tan and Quoc Le. “Efficientnet: Rethinking model scaling for convolutional neural networks”. In: *Proceedings of the International conference on machine learning*. 2019, pp. 6105–6114.

[185] Vajira Thambawita et al. “The Medico-Task 2018: Disease Detection in the Gastrointestinal Tract using Global Features and Deep Learning”. In: *In Proceedings of the CEUR Worksh. Multim. Bench. Worksh. (MediaEval)*. 2018.

[186] Jonathan Ho, Ajay Jain, and Pieter Abbeel. “Denoising diffusion probabilistic models”. In: *Advances in neural information processing systems* 33 (2020), pp. 6840–6851.

[187] Prafulla Dhariwal and Alexander Nichol. “Diffusion models beat gans on image synthesis”. In: *Advances in neural information processing systems* 34 (2021), pp. 8780–8794.

[188] Muzaffer Özbey, Onat Dalmaz, Salman UH Dar, Hasan A Bedel, Şaban Özturk, Alper Güngör, and Tolga Çukur. “Unsupervised medical image translation with adversarial diffusion models”. In: *IEEE Transactions on Medical Imaging* (2023).

[189] Cheng Peng, Pengfei Guo, S Kevin Zhou, Vishal M Patel, and Rama Chellappa. “Towards performant and reliable undersampled MR reconstruction via diffusion model sampling”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2022, pp. 623–633.

[190] Zolnamar Dorjsembe, Sodtavilan Odonchimed, and Furen Xiao. “Three-dimensional medical image synthesis with denoising diffusion probabilistic models”. In: *Medical Imaging with Deep Learning*. 2022.

[191] Afshin Bozorgpour, Yousef Sadegheih, Amirhossein Kazerouni, Reza Azad, and Dorit Merhof. “Dermosegdiff: A boundary-aware segmentation diffusion model for skin lesion delineation”. In: *International Workshop on PRedictive Intelligence In MEdicine*. Springer. 2023, pp. 146–158.

[192] Yijun Yang, Huazhu Fu, Angelica I Aviles-Rivero, Carola-Bibiane Schönlieb, and Lei Zhu. “Diffmic: Dual-guidance diffusion network for medical image classification”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 95–105.

[193] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. “Adding conditional control to text-to-image diffusion models”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 3836–3847.

[194] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. "High-resolution image synthesis with latent diffusion models". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 10684–10695.

[195] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. "Learning transferable visual models from natural language supervision". In: *Proceedings of the International Conference on Machine Learning (ICML)*. 2021, pp. 8748–8763.

[196] August DuMont Schütte, Jürgen Hetzel, Sergios Gatidis, Tobias Hepp, Benedikt Dietz, Stefan Bauer, and Patrick Schwab. "Overcoming barriers to data sharing with medical image generation: a comprehensive evaluation". In: *NPJ digital medicine* 4.1 (2021), p. 141.

[197] Maria JM Chuquicusma, Sarfaraz Hussein, Jeremy Burt, and Ulas Bagci. "How to fool radiologists with generative adversarial networks? A visual turing test for lung cancer diagnosis". In: *Proceedings of the 15th international symposium on biomedical imaging (ISBI 2018)*. 2018, pp. 240–244.

[198] Ultralytics. *YOLOv8*. https://github.com/ultralytics/ultralytics/. Accessed: 01 Aug 2023. 2023.

[199] Ge-Peng Ji, Guobao Xiao, Yu-Cheng Chou, Deng-Ping Fan, Kai Zhao, Geng Chen, and Luc Van Gool. "Video polyp segmentation: A deep learning perspective". In: *Machine Intelligence Research* 19.6 (2022), pp. 531–549.

[200] Vanshali Sharma, Pradipta Sasmal, MK Bhuyan, and Pradip K Das. "Keyframe Selection from Colonoscopy Videos to Enhance Visualization for Polyp Detection". In: *2022 26th International Conference Information Visualisation (IV)*. IEEE. 2022, pp. 426–431.

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

[202] Debesh Jha, Nikhil Kumar Tomar, Vanshali Sharma, and Ulas Bagci. "TransNetR: Transformer-based Residual Network for Polyp Segmentation with Multi-Center Out-of-Distribution Testing". In: *MIDL* (2023).

[203] Hayato Itoh, Masashi Misawa, Yuichi Mori, Masahiro Oda, Shin-Ei Kudo, and Kensaku Mori. *SUN Colonoscopy Video Database*. http://amed8k.sundatabase.org/. 2020.

[204] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. "Demystifying mmd gans". In: *arXiv preprint arXiv:1801.01401* (2018).

Brief Biography of the Author

Vanshali Sharma joined the Ph.D. program at the Department of Computer Science and Engineering of the Indian Institute of Technology Guwahati (IITG), India, in January 2019. She also worked in collaboration with Northwestern University, Chicago, USA and Chubu University, Kasugai, Japan. Before joining the Ph.D. program, she worked as a senior research fellow in a DST project at ICAR-Indian Agricultural Research Institute, New Delhi, India. She received her Master of Technology degree from the Indian Institute of Technology (ISM) Dhanbad, India, in May 2018 and her Bachelor of Technology degree from UIET, Maharshi Dayanand University, Rohtak, Haryana, India in May 2015. During her Ph.D., her research proposal was selected for the DST-INSPIRE fellowship. She was also awarded with AAAI Travel Grant 2023. Other than conference presentations, she was selected to attend Research Week with Google 2023 in Bangalore, India. Her research interest includes Computer Vision, Medical Image Analysis, and Deep Learning.

Contact Information

Email	:	vanshalisharma@iitg.ac.in, sharmavanshali@gmail.com
Web	:	https://sites.google.com/view/vanshalisharma/