AUTOMATED DETECTION AND CLASSIFICATION OF POLYPS IN COLONOSCOPY VIDEOS

PRADIPTA SASMAL

AUTOMATED DETECTION AND CLASSIFICATION OF POLYPS IN COLONOSCOPY VIDEOS

A

Thesis submitted

for the award of the degree of

DOCTOR OF PHILOSOPHY

By

PRADIPTA SASMAL

DEPARTMENT OF ELECTRONICS AND ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

GUWAHATI - 781039, ASSAM, INDIA

APRIL 2022

Certificate

This is to certify that the thesis entitled "**AUTOMATED DETECTION AND CLASSIFICATION OF POLYPS IN COLONOSCOPY VIDEOS**", submitted by **Pradipta Sasmal** (156102005), a research scholar in the *Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati*, for the award of the degree of **Doctor of Philosophy**, is a record of an original research work carried out by him under my supervision and guidance. The thesis has fulfilled all requirements as per the regulations of the institute and in my opinion has reached the standard needed for submission. The results embodied in this thesis have not been submitted to any other University or Institute for the award of any degree or diploma.

Dated:

Guwahati.

Prof. M.K. Bhuyan
Professor
Dept. of Electronics and Electrical Engg.
Indian Institute of Technology Guwahati
Guwahati - 781039, Assam, India.

To

The Almighty Lord Shiva

for His blessings

My guide **Prof. M.K. Bhuyan**

for his guidance and inspiration

&

My parents

Mr. Birendra Kumar Sasmal & Mrs. Puspanjali Sasmal

for their blessings and support

&

My wife **Suchismita**

for her love and sacrifice

Acknowledgements

I am obliged to GOD for his divine guidance and blessings. I dedicate my thesis to lord Shiva and my family. This thesis would not have been possible without several people's immense help and support in various measures. I would like to convey my acknowledgment to all of them.

First and foremost, I express my sincere gratitude to my research supervisor, Prof. M.K. Bhuyan, for providing me with an opportunity to work under his guidance. It is very difficult to describe my feelings in words to acknowledge my supervisor for his continuous guidance, constant motivation, and support throughout the doctoral studies. I am very much thankful to him for helping me with my research and other issues. It would be impossible for me to bring the research and thesis to this form without the ample facilities he provided in the IPCV laboratory and the freedom to work independently.

I am thankful to my doctoral committee members, Dr. P. Guha, Dr. S. Sundaram, and Dr. A. Anand, for their encouragement and valuable suggestions on my work. I would like to thank faculty members and the office staff of the Department of Electronics and Electrical Engineering, IIT Guwahati, for their help in carrying out this research work. I sincerely thank Prof. Y. Iwahori for his valuable suggestions on my research work. I would like to thank him for accepting me as an intern to work with him at Chubu University, Japan. I would also like to thank Prof. Kunio Kasugai of Aichi Medical University Hospital, Japan, for providing materials useful for my research work. I am thankful to my friends Tilendra Choudhary, Debajit Sarma, Shikha Baghel, Vinneta Das, E Prabhakararao, and Vanshali Sharma for their assistance in writing my thesis. I am thankful to Aniruddha Mazumdar, Nayan Moni Baisya, Rijjuban Rangslang, Saquib Mazhar, Pallab Jyoti Dutta, Allen Pattnaik, Avinash Paul, and all other members of the IPCV Laboratory. I am thankful to my wife, Suchismita, for her sacrifice and support. She stood by me and helped me keep solid and patient during the research journey. I am also thankful to my younger brother for his support.

I attribute this achievement to my parents and parents-in-law for their constant blessings, support, and silent prayers for my success and making me stand in this position.

Pradipta Sasmal

Abstract

Continuous assessment for an early and accurate diagnosis of colorectal cancer (CRC) is most important for better prognosis and clinical management. CRC is considered one of the leading causes of death worldwide. Polyps are the precursor to such cancer. Colonoscopy is the medical screening modality used to detect polyps in the mucosa of the colon. Firstly, the doctors detect and localize the polyps using the captured colonoscopy video frames. Then the polyps are resected, i.e., the region of interest (ROIs) are segmented out from the normal mucosa. Subsequently, useful features of the polyps are analyzed for dysplasia grading. Therefore, a typical colonoscopy procedure includes polyp detection and localization, segmentation, and classification. However, manual inspection and annotation of the polyps are cumbersome and inefficient due to similar pathological manifestations of the diseases on the hugely acquired frames. The polyp features may not be visible to the naked eye, making it difficult to diagnose the abnormalities. Therefore, Our current work proposes automated and efficient frameworks for polyp analysis using the colonoscopy video frames. The proposed approaches can do a virtual biopsy in detecting dysplasia in polyps. These may help lessen the burden on the clinicians, provide early and quick diagnosis, better decision-making, and telemonitoring. The automated diagnostic systems proposed by our methods can be adopted in the medical setup to diagnose diseases in polyps effectively.

In the proposed first method, important polyp cues like color, shape, and texture are incorporated into the modified particle filtering framework to track and localize the polyps in each frame of the colonoscopy video. Subsequently, the polyps are segmented using active contour (AC). This method can handle specularity and occlusion, which are generally encountered during colonoscopy. Our second approach simultaneously detects and localizes the polyps in the colonoscopy videos for real-time analysis. The proposed method is based on a deep learning-based attention YOLOv4 architecture. The proposed spatial and channel attention blocks are incorporated into the YOLOv4 framework. Results

suggest that the the proposed model performs better than the state-of-the-art methods, and the generalization and robustness of this method could be validated. The localized polyps are further classified into malignant (cancerous/adenomatous) and benign (non-cancerous/hyperplastic). Delineation or segmentation of the polyps leads to 3-D visualization, better resection, and classification. As a preliminary work, clinically significant frames are extracted using the depth information of the polyps, followed by their segmentation. However, this approach cannot be applied to flat and serrated polyps. Therefore, we proposed two segmentation approaches utilizing the dominant polyp cues for different polyp structures. The first approach is based on an unsupervised adaptive Markov-random field (MRF), which encapsulates the polyp's global texture and spatial information. Most of the existing methods in this domain are supervised. Our approach provides a competitive polyp segmentation performance while the requirement for massive labeled data is avoided. The third approach uses a saliency map guided geometric shape compactness prior for better polyp segmentation. Textural information and shape information is used for polyp segmentation. For the classification of the polyps, three methods are proposed. The first method uses the local texture and the polyps' shape information, which the doctors generally study for cancer detection. The second method combines shape features and the deep embedded features learned via a deep siamese network. In this work, poly ROIs obtained using our proposed attention YOLOv4 are used. The proposed spatial shape descriptor, pyramid histogram of oriented gradients (PHOG) features, extracts the local shape and spatial layout information of polyp images of each class. In contrast, the embedded features extract the discriminating features from each category's samples. Sometimes, the clinicians analyze the histopathology of the polyps for grading of cancer. In this view, we proposed a semi-supervised approach based on a generative adversarial network (GAN) for CRC grading using histopathological images. Experimental results validate the proposed method's efficiency even in a minimal data environment.

Keywords: Colorectal cancer (CRC), polyps, attention YOLOv4, active contour (AC), Markov-random field (MRF), modified particle filtering, generative adversarial network (GAN).

Contents

List of Figures

List of Tables

List of Acronyms

3D-FCN	3 Dimensional Fully Convolutional Network
ADR	Adenoma Detection Rate
AFLS	Adaptive Fuzzy Logic System
AC	Active Contour
AR	Augmented Reality
AP	Average Precision
ADMM	Alternating Direction of Multipliers
ANOVA	Analysis of Variance
BseNet	Binary size Estimation Network
BoW	Bag of Words
CRC	Colorectal Cancer
CT	Computed Tomography
CCD	Charge-coupled Device
CRF	Conditional Random Fields
CNN	Convolutional Neural Network
cGAN	Conditional Generative Adversarial Network
C3dNet	Convolutional 3-dimensional Network
CWC	Color Wavelet Covariance
CV-LSM	Chan-Vese Level Set Method
DAS	Diagnostic Assistance System
DT	Decision Tree
DCF	Discriminative Correlation Filter
DCT	Discrete Cosine Transform

DSC	Dice Similarity Coefficient
DBD	Differential Box-Counting
EVAR	Energy Variance
EM	Expectation Maximization
FCN	Fully Connected Neural Network
FICE	Flexible Imaging Color Enhancement
FPS	Frames Per Second
FD	Fractal Dimension
FWLBP	Fractal Weighted Local Binary Pattern
GI	Gastrointestinal
GLCM	Gray Level Co-occurance Matrix
GAN	Generative Adversarial Network
GUI	Graphical User Interface
GRF	Gibbs Random Field
HD	High Definition
HOG	Histogram of Oriented Gradient
HD	Hausdorff Distance
IoU	Intersection over Union
ICM	Iterative Condition Mode
K-NN	K-Nearest Neighbors
LCI	Lined Color Imaging
LBP	local Binary Pattern
LDA	Linear Discriminant Analysis
LSTM	Long Short-Term Memory
MLP	Multi-Layer Perceptron
MDE	Monocular Depth Estimation
MRF	Markov Random Field
MAP	Maximum *a Posteriori* Probability
MI	Mutual Information
NBI	Narrow Band Imaging

NN	Neural Networks
ORB	Oriented FAST and Rotated BRIEF
PCA	principal Component Analysis
PCP	Principal Component Pursuit
PCT	Principal Component Transform
PSO	Particle Swarm Optimization
PHOG	Pyramid Histogram of Oriented Gradient
ROI	Region of Interest
RF	Random Forest
SIFT	Scale-Invariant Feature Transform
SFS	Sequential Forward Selection
SPM	Spatial Pyramid Matching
SSD	Single Shot MultiBox Detector
SDSP	Saliency Detection by combining Simple Priors
SLIC	Simple Linear Iterative Clustering
SSL	Semi-Supervised Learning
TA	Tubular Adenoma
TVA	Tubulu-Villous Adenoma
TLD	Tracking Learning Detection
V-GAN	Variational Generative Adversarial Network
VR	Virtual Reality
VAR	Variance
WCE	Wireless Capsule Endoscopy
WL	White Light

List of Symbols

H	Mutual entropy
w_k	weight of particle
k	scaling constant
$\vec{X_k}$	State of the tracker
$\vec{\mathcal{N}}$	measurement of noise
$G(\mathbf{u})$	Transfer function of log Gabor filter
$\mathbf{u}$	coordinates in the frequency domain
ω_0	center frequency of the Gabor filter
σ_F^2	variance about the center frequency
$f_L(\mathbf{x})$	Filtered output of L channel of CIEL*a*b color space
$f_a(\mathbf{x})$	Filtered output of a channel of CIEL*a*b color space
$f_b(\mathbf{x})$	Filtered output of b channel of CIEL*a*b color space
maxa	Maximum value of $f_a(\mathbf{x})$
mina	Minimum value of $f_a(\mathbf{x})$
maxb	Maximum value of$f_b(\mathbf{x})$
minb	Minimum value of $f_a(\mathbf{x})$
$S_F(\mathbf{x})$	Saliency map for frequency prior
$S_C(\mathbf{x})$	Saliency map for color prior
$S_D(\mathbf{x})$	Saliency map for location prior
$S_S(\mathbf{x})$	Saliency map for feedback prior
$V(\mathbf{x})$	Final saliency map
H	Height of a feature map
W	Width of a feature map

C	Depth of a feature map
M	Input feature map
I_c	Channel descriptor
$\mathbf{d}$	Computed inverse depth
$\mathbf{d}'$	Ground truth inverse depth
N	number of pixels in a frame
s	Scale
t	shift
$\mathcal{L}$	Loss function
$\mathcal{L}_r()$	Regularization term
Q^k	Difference of inverse depth maps at a scale k
d	Moment distance
ΔS	Gradient Magnitude
w_i	weight of the normalized score
f	Fused score
Ψ	Set of image lattice sites
x	Pixel value
ζ	Clique
N_ψ	neighborhood system
Z	Normalization constant
T	Temperature parameter
$U(x)$	Energy function
$V_c(x)$	potential function
$F = f$	Features extracted from an image I F is a random variable, and f is an instance of it
$\Phi(f_\psi \mid Y = y)$	Data penalty term which penalizes a pixel ψ with a label y for given features f
$\theta_{\psi,\ell}(y_\psi, y_\ell)$	Penalty term used to maintain the smoothness of the label field
β	A constant in MRF formulation
μ_m^l	Mean for l^{th} feature component belonging to m^{th} class-label
σ_m^l	standard deviation for l^{th} feature component belonging to m^{th} class-label
s_x	features of x^{th} superpixel

$dist\|\|\cdot\|\|$	Euclidean distance
E_f	Feature modeling component
E_l	Region labeling component
R	Radius in LBP calculation
S	Shape compactness
Λ	2-Dimensional image domain
E_{up}	Unary potential prior
E_{cs}	Shape compactness prior
$A(\mathbf{x})$	Area in digital domain
$\beta(ji)$	Pairwise potential term
L_m	Laplacian matrix
$\mathbf{1}$	A vector with value one for each element
χ	Chi-square distance
L	Pyramid resolution level
B	A bounded set in an n-dimensional euclidean space S
$D(b)$	A density function represents fractal dimension
N_r	Non-ovelarring copies of 2-D structures
r	Variance of Gaussian scale-space with
$\Lambda_{x,y,r}$	Gaussian scale-space
$\Gamma(x,y,r)$	Intermediate image matrix of FD
μ_{ik}	Fuzzy membership value of k^{th} vactor in i^{th} class
m	Fuzzification parameter
f_i^a	Anchor embedding
f_i^p	Positive embedding
f_i^n	Negative embedding
G	Generator
D	Discriminator
z	Noise vector
$p_{data}(x)$	Data distribution
x_g	Generated images

x_r	Real images
G_m	Modified generator
D_m	Modified discriminator
L_{sup}	Supervised discriminator loss
L_{unsup}	Unsupervised discriminator losses

1

Introduction

Contents

Objective

This chapter describes the importance of early and accurate diagnosis of colorectal cancer (CRC), which is considered one of the leading causes of death worldwide. The need for an automatic diagnostic assistance system (DAS) for cancer detection using colonoscopy videos is highlighted. This chapter discusses the existing polyp analysis frameworks and motivates solutions for achieving better performances. The framework's objectives are automated detection, localization, segmentation, and classification of polyps found during a colonoscopy. Polyps are defined as the overgrowth of tissue generally found in the mucosa of the colon region and are the precursor to cancer. Early diagnosis of such cancer is crucial for better prognosis and clinical management. Doctors generally detect and segment the polyps from the mucosa and do a comprehensive study on them for cancer detection. Therefore, this chapter discusses the strategies to automate these processes with near-accurate diagnosis. Existing works on these tasks using colonoscopy videos are reviewed, and the advantages and limitations of these methods are elaborated. Primarily, the source of motivation originates from the fact that the polyp region's image characteristics are different from the non-polyp area. Similarly, benign (non-cancer) polyp features differ from cancer (malignant) polyp features. Therefore, effective feature learning holds the key to developing an effective DAS. The chapter then defines the research issues to be addressed for the development of a DAS for each of the tasks. The chapter concludes with a brief description of the organization of the thesis.

1.1 Colon Cancer

Colorectal cancer (CRC) is one of the major health crises globally. The high mortality of CRC contributes significantly to the total deaths, and it is considered to be the third most frequently occurring cancer [1]. It is the fourth leading cause of death worldwide (9.4%) [2]. The global CRC is estimated to reach 2.2 million new cases and 1.1 million deaths by 2030 [3]. Such cancer in its initial state is called polyp and is generally benign. Polyps are abnormal tissues and are usually found in the mucosa of the colon [4]. Therefore, improvement in early diagnosis, screening, and pre-clinical intervention of colonic lesions would help in better prognosis and clinical management of the disease. When colorectal cancer is found early, the five-year survival rate rises to 90 percent, compared to 15 percent when it is detected at an advanced stage [1]. Figure 1.1 shows the 5-year relative survival rates

[1] https://seer.cancer.gov/statfacts/html/colorect.html

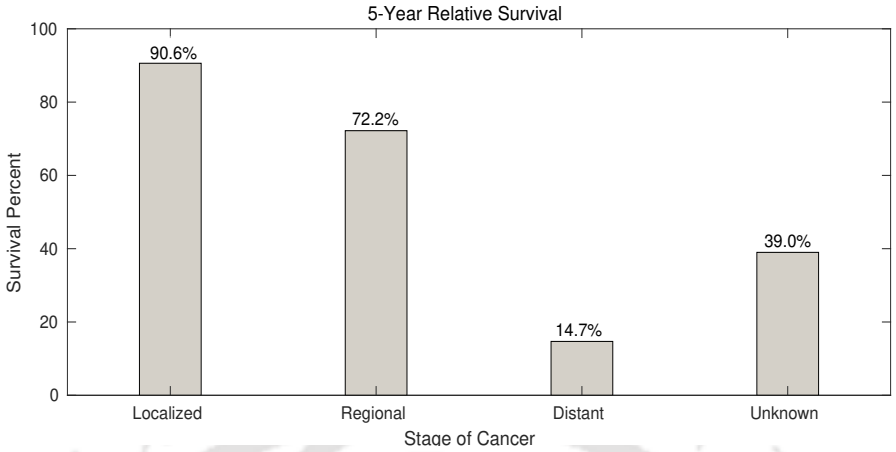


Figure 1.1: Five-year survival rates of the colorectal cancer for each stage. (a) localized (stage I), i.e., confined to its initial site, (b) regional (stage II), i.e., spread to nearby lymph nodules, (c) distant (stage III), i.e., metastatis happened, (d) unknown (unstaged).

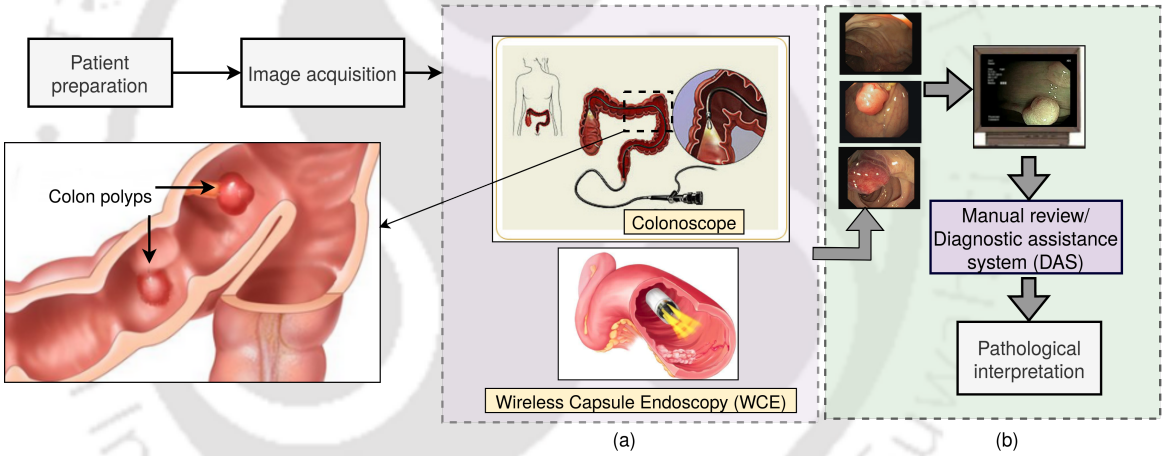


Figure 1.2: Acquisition of colonic video sequences for polyp detection. (a) different screening modalities, and (b) analysis of colonic video frames.

for cancer. From Figure 1.1, it can be seen that the rate of survival increases with an early diagnosis. Therefore, regular and timely screening of the colon region can significantly reduce the risk of cancer substantially [3].

Endoscopy is the modality used in medical setup for screening the entire gastrointestinal (GI) tract for polyp detection. During endoscopy, the clinician inserts the cable (scope), having a flexible camera and light to capture the internal images of the GI tract. There are two types of endoscopy in practice: upper endoscopy and lower endoscopy. The lower endoscopy can further be classified in categories like sigmoidoscopy, colonoscopy, etc., based on the portion of the digestive tract under scrutiny. Since colon cancer is the most frequently occurring and prevalent cancer, our current study is limited to

analysis of colon cancer only. Colonoscopy is considered the gold standard in colon cancer screening [5]. The endoscopist navigates the scope to find abnormalities like bleeding, ulcer, polyp, etc., in the colon region. Wireless Capsule Endoscopy (WCE) is another modality to monitor the conditions in the tract. It is a capsule-like device swallowed by the patient, and images are recorded and transmitted wirelessly to a nearby recording system. During the colonoscopy, doctors comprehensively analyze the detected polyp regions to find dysplasia in them. Depending on the condition of the polyp nature, they may opt for laparoscopic surgery. However, the number of frames captured during the entire colonoscopy process is so humongous that it challenges the surgeon to infer useful clinical information. Therefore, video summarization techniques are adopted that only retain the clinically informative frames. Figure 1.2 illustrates the general pipeline for colonoscopy image analysis. Other non-invasive modalities for colon screening include computed tomography (CT) and color Doppler ultrasound. The CT, otherwise called the virtual colonoscopy, produces 2-D or 3-D reconstructed images of the colon. Though the method takes very little time to detect lesions, the usual colonoscopy procedure must be done for biopsy and resection of such detected lesions. Therefore, colonoscopy is the most adopted procedure in a clinical setup for anomaly detection in the colon. Abnormalities like polyps may be neoplastic or non-neoplastic. Non-neoplastic polyps are normal tissues, whereas neoplastic or adenomatous polyps are cancerous [6]. These polyps are over-growth of tissues and look like tumors. These polyps become cancerous if left untreated. Ideally, the endoscopists review the captured colonoscopy frames to detect them. After finding the frames having polyps, the doctors extract the Region of interest (ROI), i.e., the polyp regions. Then, they analyze different characteristics of these regions, such as geometry (shape and size), the surface of the polyp (texture), color (blood traces), boundary (smooth or wavy), etc., to classify them into different classes. However, many constraints incurred during polyp analysis make the diagnosis process inefficient and cumbersome.

1.2 Polyps

Neoplastic polyps result from abnormal cell proliferation and are benign when they are confined to the mucosa. These lesions become malignant when they start invading the submucosa. Subsequently, they spread through the lymph and circulatory systems, causing metastasis. The size and shape of these polyps are related to the degree of severity of cancer. Smaller polyps (diminutive) are generally benign, whereas big polyps tend to have malignancy properties [7]. Different classification schemes

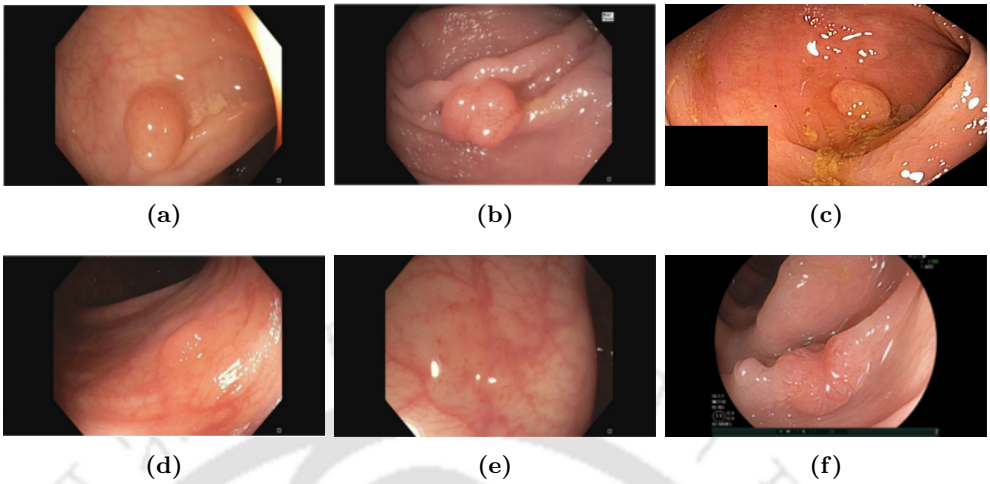
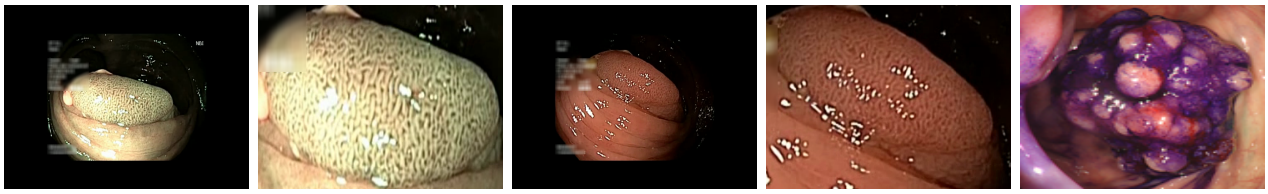


Figure 1.3: Paris classification of polyps. (a) pedunculate (0-Ip), (b) subpedunculate (0-Isp), (c) sessile (0-Is), (d) slightly elevated (0-IIa), (e) completely flat (0-IIb), (f) slightly depressed (0-IIc).

based on polyp morphological features, like pit pattern, vessel pattern, surface pattern, color and shape, and size, are available in the literature [8]. The Paris classification [8] gives a framework for the classification of colonic polyps, which is well accepted. Figure 1.3 shows some representative images based on Paris classification. Pedunculated and sessile polyps are easier to detect than flat polyps. Other types of polyps are difficult to discriminate from the normal tissues through naked eyes. In such cases, there is a need for a computer-aided diagnostic system.

Imaging technologies for polyp detection

Different imaging technologies for polyp detection have been proposed over time. Earlier, chromoendoscopy was used extensively to detect polyps in the colon. Dyes such as methylene blue or indigo carmine are applied to distinguish polyp from normal tissues. Optical technologies such as narrow-band imaging (NBI) (Olympus), flexible spectral imaging color enhancement (FICE) (Fujinon), and i-scan (Pentax) are nowadays used. Dye-based colonoscopies are effective for detecting long-term inflammation. Optical imaging techniques are effective against diminutive polyp detection. White-light (WL) colonoscopy uses the full spectrum (wavelength: 400-700 nm), whereas NBI uses the narrow-band blue-green light, which improves visualization of the neoplastic lesions against the adjacent normal mucosa. The FICE and i-scan systems process reflected photons to reconstruct virtual images of the polyp. Figure 1.4 shows representative samples acquired with different imaging techniques [9]. WL and NBI techniques are mostly used during the colonoscopy procedure. The publicly available



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

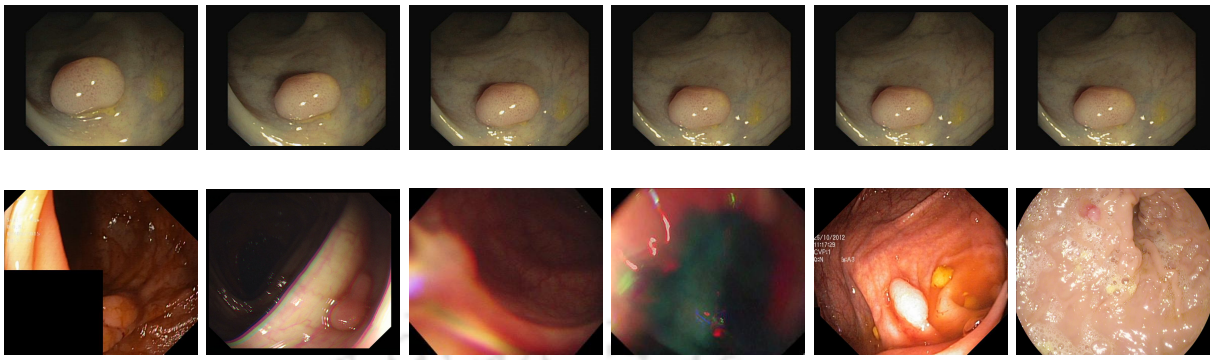
Figure 1.4: Colonoscopy image acquisition using different imaging modalities. (a) NBI, (b) corresponding ROI, (c) WL, (d) corresponding ROI, (E) Dye.

colonoscopy images and videos datasets are mostly acquired using WL and NBI modalities.

1.3 Challenges in Polyp Analysis using Colonoscopy

During the regular colonoscopy, the CCD sensors and the camera are under the influence of involuntary muscle movement. In WCE, on the other hand, the capsule moves under the peristalsis movement, and it is challenging to control the motion and orientation of the camera. Thus, redundant and clinically non-significant frames are generally obtained in a video sequence. WCE takes nearly 8 hours, capturing close to 50000 frames. A large part of the data is clinically not relevant and needs to be removed [10]. The captured frames are, therefore, susceptible to some of the degradation factors discussed below:

- Black mask and text: The black mask formed around the frame conveys no clinical information. The text that gives the test information may lead to the hindrance of polyp, which may lie along the border of the mask.
- Ghost colors: The improper synchronization of color channels may lead to ghost colors. This phenomenon degrades the quality of captured frames.
- Motion blur: The colonoscopy frames are often susceptible to motion blur and produce a hazy and blurred image.
- Specularity: Specular highlights are saturated portions of an image. It hinders the useful contextual information of an image's ROI.
- Low illumination: The considerable variation of lighting is expected in the colonoscopy procedure. This may add to the polyp miss detection rate.



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

Figure 1.5: Issues in manual analysis of colonoscopy frames. First row frames are consecutive frames of a video sequence conveying the same information, (a) polyp is obstructed by mask, (b) ghost colors, (c) motion blur, (d) low illumination, (e) specular highlights, (f) waste materials and bubbles.

- Other tissues and materials: The presence of water bubbles, waste materials, normal tissues mimicking the abnormality, etc., may reduce the efficiency of polyp analysis.

An effective clinical interpretation of the polyp features becomes very difficult with the presence of such factors. The representative images illustrating such elements are shown in Figure 1.5. Image frames shown in the first row of Figure 1.5 belong to some consecutive frames recording of a patient. It can be seen that the frames contain almost similar information. Generally, in colonoscopy, redundant and duplicate frames are often captured.

Therefore, reviewing each frame manually for polyp detection from a hugely acquired colonic frame is very difficult and inefficient. The features of polyps are so indistinctive that sometimes it is challenging to distinguish them from the normal colon tissues. Also, the maximum polyp detection rate, which can be achieved through colonoscopy, is less than 50 % as it is highly operator dependent [11]. Therefore, it is essential to reduce the miss detection rate of polyps. However, manual inspection and annotation of polyps are cumbersome and inefficient due to the big medical images acquired through colonoscopy and similar pathological manifestations across diseases. Sometimes, the polyp features may not be visible to the naked eye, and also, the diagnostic information is sometimes overlooked, which makes it very difficult in decision making.

The application of new technologies in health care applications is on a constant rise. With the advent of new modalities, efforts have been made to enhance the efficiency of colonoscopy. Recently, optical endoscopic modalities using narrow-band imaging (NBI) have been developed for improved

colorectal lesions detection [12, 13]. NBI imaging enhances the lesion's vascular pattern, thereby increasing their discriminating ability. Blue light imaging and i-Scan endoscopy are also used for better polyp detection [14,15]. Another endoscopy modality that acquires high definition (HD) images is linked color imaging (LCI). This imaging technique uses color as an important cue for lesions. Generally, the color of a malignant (Adenomatous) polyp looks reddish, and the color of a non-adenomatous polyp is whitish [16]. LCI enhances the red and white color and makes the red area look more reddish, and the white area looks more whitish during colonoscopy. Thus, LCI helps in lesion detection and helps in their classification. Other techniques adopted to improve lesion detection include better bowel preparation, use of the broad field of view camera, flattening of colonic folds, etc. [17]. However, the diagnosis using these techniques for better polyp detection during colonoscopy needs a highly experienced and trained expert in this domain [18]. Another problem that arises during lesion detection in colonoscopy frames is the high variability in the polyp characteristics. Typically, small or serrated polyps, diminutive and isochromatic flat polyps, are missed during manual inspection. Also, device and patient-specific colonoscopic frames will have different image characteristics [19].

1.4 Diagnostic Assistance System (DAS) for Colonoscopy Image Analysis

To adress the above mentioned challenges, automated diagnostic systems have been adopted and investigated for colonoscopy images. These methods have shown effectiveness in the pathological interpretation of diseases. Our current work focuses on dysplasia grading of the polyps detected during colonoscopy. Therefore, the steps generally followed by an endoscopist are very crucial in determining the polyp grades. These steps are now defined as follows: As discussed earlier, the doctors try to classify the polyps after a vivid analysis of the polyps. Firstly, they detect and localize the polyps in the frames. Not all the captured frames are clinically significant. Therefore, frames having polys are retained for further analysis. After polyp detection, they try to segment the polyp, i.e., the region of interest (ROI), for extracting clinical information from them. After feature extraction, polyps are classified into various grades according to the carcinoma stages. Therefore, typical analysis of colonoscopy images includes polyp detection and localization, segmentation, and classification. Other applications include 3D visualization and reconstruction of polyp surfaces which enables doctors to have a better understanding of the polyps. The automated system does a virtual biopsy in detecting dysplasia in polyps.

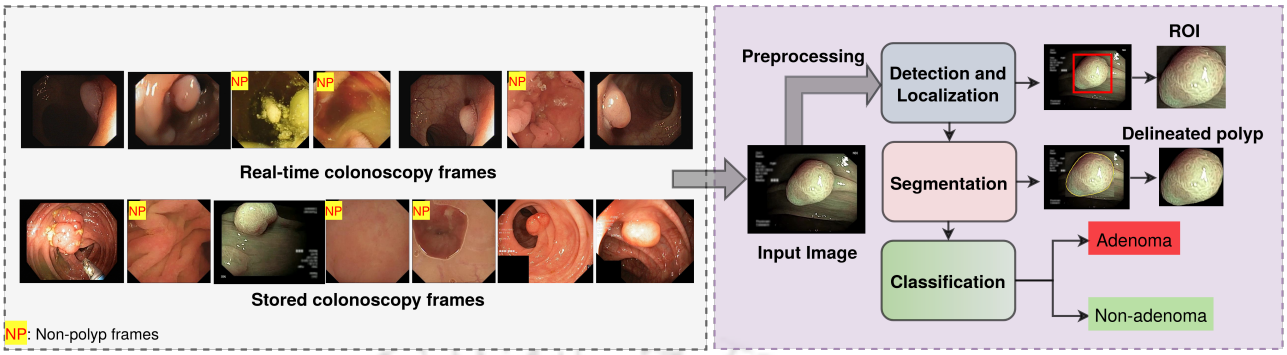


Figure 1.6: A typical colonoscopy image analysis framework.

These systems have made significant progress in medical image analysis based on machine learning and deep learning algorithms. The foundation work for the development of computer-aided diagnosis (CAD) for cancer diagnosis using chest radiography began around 1980 [20]. In the early stage, these models were based on hand-crafted feature extraction. Also, the lack of inefficient image acquisition modalities contributes to poor image analysis. These models have effectively performed image and video analysis with advancements in machine learning approaches. Deep learning model-based CADs have been extensively used in polyp detection and diagnosis of cancer [21–23]. These systems help in achieving better adenoma detection rate (ADR) and assist in polypectomy of hyperplastic polyps [22, 24, 25].

A lot of works proposed in this domain are available in the literature. But, several different definitions have been used in literature for detection, localization, segmentation, and classification. In this thesis, the meaning of the aforementioned tasks is as per the terminology used by the Gastrointestinal Image Analysis (GIANA) challenge [2]. Figure 1.6 illustrates a typical pipeline for the classification of polyps. Polyp detection is the task of identifying polyp frames from the non-polyp frames. The left block of Figure 1.5 shows some of the acquired colonoscopy video frames. The top row of this block shows some consecutive frames captured for a patient in real-time. It can be seen that some uninformative frames are also captured during image acquisition. Therefore, selecting the polyp frames out of the hugely acquired frames is vital. This step is called the polyp detection stage. Subsequently, the position of the polyp in these corresponding detected frames is localized to find the ROI. ROIs are the clinically significant regions. These ROIs can be used to extract features for polyp classification. For polyp classification, only polyp regions are sufficient. Therefore, finding the exact polyp boundary

[2] https://giana.grand-challenge.org/Tasks/

to delineate is also preferred. This task is called segmentation. It is also helpful in the resection of polyps. Finally, the malignancy in polyp is detected. Broadly, it is classified as adenoma (malignant) and non-adenoma (benign). Further, polyp grading like Tubular adenoma (TA)- TA low grade and TA high grade, Tubulu-Villous adenoma (TVA)- TVA low grade and TVA high grade, hyperplastic (benign), etc., are also found in the literature [26]. In this thesis, our analysis is based on an automated polyp classification design. It is to be noted that all the intermediate steps are vital for the final polyp classification.

1.5 Colonoscopy Image Analysis: A Review

Much work on polyp analysis using colonoscopy images can be found in the literature. The critical stages shown in Figure 1.6 are discussed with different degrees of exploration. A particular proposed method may address any of the steps or steps in integration to achieve an accurate diagnosis of polyps. Since our thesis explores to address all of them, a systematic review of each task is important. Firstly, we will review the existing polyp detection approaches available in the literature. These proposed techniques for all these tasks are elucidated in chronological order.

1.5.1 Polyp detection in colonoscopy images

Colour, texture, and shape are essential cues for polyp detection. These features have been extensively used in polyp characterization by hand-crafted based methods. Foundation works on polyp detection are on CT colonography or virtual colonoscopy images. Most of the research works are based on geometric features for polyp detection [27–30]. Karkanis et al. [31] used color wavelet features as the texture features to detect polyps. However, the performance in terms of sensitivity falls for polyps having little texture. One of the earliest works on automatic detection of polyps was done by Kodagiannis et al. [32]. They extracted color texture features by utilizing spectrum from six channels (red-green-blue: RGB and hue-saturation-value: HSV color spaces) with an adaptive fuzzy logic system (AFLS) as the classifier. They achieved 97% sensitivity on a relatively very small dataset of 140 images with 70 polyp frames. They assumed that different color spaces better represent color texture patterns to characterize the polyp features. However, it is known that the color channels in RGB images tend to exhibit high levels of correlation. Similarly, only color features may not be sufficient to characterize the polyps as the non-polyp regions also show similar color characteristics. Li et al. [33] used two shape descriptors, MPEG-7 shape descriptor (angular radial transform - ART) and Zernike moments.

They achieved an accuracy of 86.1% on 300 test images, out of which 150 contained polyps. However, non-polyp tissues having a similar polyp shape can be found in the colonoscopy frames. Later, they utilized the color and shape features to obtain a better polyp detection performance [34]. In [35], a Log-Gabor filter-based segmentation along with SUSAN edge detector and active contour (AC) was proposed to identify polyp candidates. However, the method assumes the shape of polyps as an ellipse, which is not the case as polyps can have different shapes and sizes. Also, a limited number of images were tested to validate their method. Hwang and Celebi [36] used Gabor texture features in a watershed segmentation framework to indicate polyp regions in a frame. Their method also assumes the shape of the polyp to be an ellipse or circle. Nawarathna et al. [37] extracted the texture features using Schmid filter bank and classified the polyp regions using SVM and K-nearest neighbors (K-NN). Later, they added the local binary pattern (LBP) feature and Leung-Malik filter for efficient texture feature extraction [38]. However, their models rely on texture features and do not consider other important polyp features. A geometry-based approach based on protrusion measures for polyp detection was proposed by Figueiredo et al. [39]. However, their model failed to detect the flat and sessile polyps as such polys do not have protrusions. Other protrusion-based approaches for polyp detection were proposed in [40–42]. Zhao et al. [43] proposed to use the temporal information along with color, edge, and texture features for polyp detection. Li and Meng [44] extracted the uniform LBP features from the image sub-patches in the wavelet transformed domain. However, their model failed to characterize the flat textureless polyps. Yuan and Meng [45] proposed a bag of features representation of polyps using SIFT and K-means clustering. Later, Yuan et al. [46] added other texture features, like LBP, uniform LBP, complete LBP, and HOG features with the SIFT feature for better polyp characterization. However, selecting patch size around the key points and finding the best number of visual words for optimum performance is difficult to obtain. The proposed method only gives the best results for a patch size of 8×8 and a visual word size of 120. Wang et al. [47], used edge-cross-section visual features to track the edges along the contour of polyps from consecutive video frames. However, their model failed to find the edges for hidden and occluded polyps in frames. The commonality among these techniques is feature extraction, which extracts discriminative features (e.g., texture, shape, color) and machine-learning classification to identify polyp regions. All the methods are summarized in Table 1.1 and Table 1.2.

Most of the techniques discussed above are based on supervised learning. However, these methods

Table 1.1: Overview of handcrafted feature learning-based polyp detection and localization techniques in colonoscopy. Abbreviations: ART—angular radial transform; BoW—bag of words; BoF—bag of features; HMM—hidden Markov model; SIFT—scale invariant feature transform; HoG—histogram of oriented gradients; LLE—locally linear embedding; LBP—local binary patterns; MLP—multilayer perceptron; SVM—support vector machines; LDA—linear discriminant analysis; NN—neural networks; CRF—conditional random fields; DT—decision tree; RF—random forest; * —no classifier used; **—not available; (*)—number of polyp images.

Method	Features/Technique	Feature type	Classifier(s)	Dataset Contents
[48]	Curvature analysis	Shape	*	6(4)
[31]	Color wavelet	Texture	NN	8
[49]	LBP	Texture	NN	3
[50]	Energy-angular second moment, correlation, inverse difference moment, entropy	Texture, color	NN	4
[51]	Area, color and shape of segments	Shape, color	*	**
[31]	Color wavelet	Texture	LDA, SVM	1380
[52]	Texture spectrum histogram, chromatic and achromatic histograms	Texture, color	SVM	66(54)
[53]	Gray level co-occurrence matrix	Texture	NN	4
[53]	Discrete wavelet transform coefficients, histogtam of CIE-Lab color space	Texture, color	Ensemble SVM, LDA	58(46)
[54]	Morphological watershed segmentation	Color, shape	*	100(50)
[55]	LBP, wavelet energy, opponent color local binary pattern	Texture, color	SVM	15,000
[56]	Curvature, RGB color	Shape, color	SVM	8621(815)
[57]	Curvature, RGB pixel color	Color	SVM	35
[58]	Co-occurrence matrix, color	Texture, color	SVM	74
[59]	Co-occurrence matrix, local binary pattern	Texture, color	*	1736
[60]	Number and size of edges, distance of edges from specular highlights	Texture, shape, temporal information	CRF	35 videos
[61]	Watershed segmentation, RGB color pixel	Color, shape	*	300 videos
[62]	Edge cross-section profiles (ECSP)	Texture, color, shape	DT, SVM	1513
[63]	Edge normal, edge cross-section feature	Texture, color, shape	RF	300(300)

Table 1.2: Overview of handcrafted feature learning-based polyp detection and localization techniques in colonoscopy (Contd.).

Method	Features/Technique	Feature type	Classifier(s)	Dataset Contents
[32]	Texture spectrum	Color	Neurofuzzy	140(70)
[32]	Texture spectrum	Color	Neurofuzzy	140(70)
[33]	ART descriptor, zernike moments	Shape	MLP	300(150)
[34]	Chromaticity histogram, zernike moments	Shape, color	MLP, SVM	300(150)
[35]	Log-Gabor filter, SUSAN edge detector	Shape	*	50(10)
[36]	Gabor filter, watershed segmentation	Shape	*	128(64)
[37]	Filter bank based texton histogram	Texture	K-NN, SVM	400(25)
[39]	Curvature analysis	Shape	-	1700(10)
[64]	Color moments, LBP	Color, texture	SVM	2 videos
[43]	Color, edge, texture, HMM	Color, texture	weak k-NN	400(200), 1120(560)
[44]	LBP, Wavelet	Texture	SVM	1200(600)
[65]	Color, edge, texture, HMM	Color, texture	Boosted-SVM	1200(600)
[66]	Color + Gabor filters + BoW	Color, texture	SVM	250(50)
[40]	Local polynomial approximation + geometry	Shape	SVM	3 videos(40)
[41]	HOG, color	Shape, color	MLP	30540(540)
[42]	Geometry, color, Monogenic LBP	Shape, color, texture	SVM	400(200)
[45]	Saliency, SIFT, BoF	Texture	SVM	872(436)
[67]	Gabor filter + monogenic LBP + LDA	Texture	SVM	872(436)
[68]	LBP + HOG	Texture, shape	Regression	27984 (12984)
[69]	RGB, Variance, radius	Color, shape	SVM	359
[70]	SIFT + BoF + K-means	Texture	SVM	800(400)
[46]	SIFT, complete LBP, BoF	Texture	SVM	2500(500)

provide inferior performances as features learned during the training of a supervised model may not be sufficient to generalize the unseen test datasets. The considerable variation of image features among the acquired data from different colonoscopy modalities gives unsatisfactory performances. Characterization of polyps using hand-curated features needs immense domain knowledge and expertise. The difference in the critical features like sizes, shapes, textures, colors, and orientations of these polyps [19] is a challenge. Therefore, all various challenges have to be addressed. Recent advancements in the learning paradigm show significant progress in this domain. The deep learning technique generally performs well because of its ability to extract hidden image features. Some of the deep learning-based approaches to polyp detection are discussed below.

Recently, deep learning-based automated polyp detection systems have been proposed for real-time polyp detection [71–74]. Ahmad et al. [75] gave a brief on the clinical applications of computer-aided diagnosis in colonoscopy. Convolutional neural networks (CNNs) based methods have been deployed in medical imaging for various tasks [76–80]. Shin et al. [81] proposed a transfer learning approach for polyp detection in colonoscopy. They used Inception ResNet and proposed a region-CNN for the task. Shin et al. [82] proposed a conditional generative adversarial network (cGAN) to generate synthetic colonoscopy images for improved detection performance. Lee et al. [74] employed YOLOv2 [83] for real-time polyp detection and localization. Yamada et al. [84] deployed Faster RCNN and VGG-16 to detect and localize lesions in endoscopic video frames. They achieved a real-time detection performance with a minimum polyp miss rate. Table 1.3 provides some state-of-the-art methods based on deep learning. From the above discussion, it is quite evident that hand-crafted-based methods utilize the dominant polyp features for polyp vs. non-polyp classification. The colonoscopy polyps are susceptible to geometric and photometric transformations. Also, noise and complex background make it challenging to learn the discriminating features. Therefore, the missed polyp detection rate of these methods cannot be discounted. The deep learning-based approaches provide superior performances compared to the conventional methods. However, the requirement of extensive labeled data during the training of these models is sometimes tough to achieve. Under such scenarios, the generalization and robustness of the detector cannot be guaranteed. Therefore, there are many scopes for improving the performance. The challenges and possible solutions are discussed in the corresponding chapter. The next important analysis is polyp segmentation. The existing literature in this regard is discussed below.

Table 1.3: Overview of deep learning-based polyp detection and localization techniques in colonoscopy. Abbreviations: CNN —Convolutional network; C3dNet—Convolutional 3-dimensional network; 3D-FCN —3 dimensional fully convolutional network; DCF —Discriminative correlation filter; BseNet —Binary size estimation network .

Method	Approach/Technique	Architecture	Datasets
[85]	Patch based feature extraction, SVM	AlexNet (Pre-trained)	CVC-ColonDB
[86]	Patch based feature extraction	CNN	CVC-ClinicDB, ETIS-Larib, ASU-Mayo
[87]	Bounding box based feature extraction	AlexNet	ASU-Mayo
[88]	Patch based feature extraction	AlexNet	ASU-Mayo
[89]	Colour, shape, and temporal patch based feature	CNN	ASU-Mayo
[90]	Egde map generation and classification	CNN	ASU-Mayo
[91]	Egde map, K-Means-with-Connectivity-Constraint (KMCC) segmentation	AlexNet	CVC-ColonDB, ASU-Mayo
[92]	Classification	CNN	ASU-Mayo
[93]	Spatio-temporal features	BseNet, C3dNet [94]	Private
[95]	Spatio-temporal features	Convolutional 3-dimensional network (C3dNet)	Private
[96]	Multi-stage frame classification	Cascaded Deep Decision Network (CDDN)	ISBI 2014 Challenge [89]
[97]	Bounding box	R-CNN (VGG-16)	CVC-ClinicDB, CVC-VideoClinicDB, CVC-ColonDB, CVC-EndoSceneStill
[73]	End-to-end training (Bounding box)	VGG-16, VGG-19, ResNet50	Private
[98]	Encoder-Decoder based feature learning	Y-Net	ASU-Mayo
[81]	Region proposal network (RPN) and False positive learning	Faster R-CNN (Inception ResNetv2)	CVC-ClinicDB, CVC-VideoClinicDB, ETIS-LARIB, ASU-Mayo
[99]	Spatio-temporal features	3D-FCN	ASU-Mayo
[100]	Spatial and temporal features	RYCO (ResYOLO, (DCF))	CVC-ClinicDB, ETIS-LARIB, ASU-Mayo
[72]	Bounding-box	SegNet	CVC-ClinicDB, Private
[101]	Color wavelet and patch based feature extraction using CNN	CNN	CVC-ClinicDB, ETIS-LARIB, ASU-Mayo
[102]	Bounding-box End-to-end training	YOLO	CVC-ColonDB, CVC-ClinicDB, ETIS-LARIB
[103]	Bounding-box End-to-end training	Darknet-YOLO	ASU-Mayo, Private
[71]	Encoder-decoder based residual network	ColonSegNet	Kvasir-SEG

1.5.2 Polyp segmentation in colonoscopy images

The localized polyps, i.e., the ROIs, are utilized in cancer detection. However, sometimes it is required to have the exact delineated polyps for the analysis. A not perfectly localized polyp includes a significant amount of non-clinical information, thereby inducing unnecessary features that may affect the overall efficiency. Also, perfect boundary detection can lead to an ideal resection of the polyps. Therefore, efforts have been made to do the localization perfectly. A significant amount of work has been carried out on this view. Earlier works are based on handcrafted feature learning techniques compared to the recent deep learning approaches. A systematic review of those methods reported in the literature is discussed.

Early work on the development of a fully automated decision support system for celiac disease diagnosis was proposed by Gadermayr et al. [104]. They extracted different image features from the colonoscopy frames for disease diagnosis. Figueiredo et al. [105] proposed an unsupervised approach for segmentation of colorectal polyps using a histogram-based two-phase segmentation model. They only used a small dataset of 87 frames collected using NBI to test their model. They produced good segmentation results, but their method's generalizability and robustness cannot be guaranteed. In [106], an Edge and neighborhood guidance network (ENGNet) is proposed to segment polyps. The suggested model's ability to capture enough spatial and texture information for efficient polyp segmentation is determined by the polyp's neighborhood areas. As a result, polyp regions with low light illumination and blur have poor segmentation results.

Most of the previous techniques employed for polyp segmentation are based on supervised learning. In [107], eight texture and color features are extracted from a gray level co-occurrence matrix (GLCM). Finally, an SVM is used to extract the region of interest. In [108], texture features and Color wavelet covariance (CWC) features are used for polyp non-polyp classification. Earlier, we proposed an active contour model using the Chan-Vese level set method (CV-LSM) after pre-processing the colonoscopy frames using principal component pursuit (PCP) [109]. However, the method produces low polyp segmentation performance in low illumination conditions. Bernal et al. [61] proposed a region descriptor for polyp segmentation. In [110], a shape-based approach is proposed to mark out the polyp regions. However, the approach generated unsatisfactory segmentation results in specific scenarios due to the polyps' uneven shape.

In [77], a deep learning approach based on a fully convolutional neural network (CNN) is employed

for polyp segmentation. The Jaccard index (IoU) was not evaluated to show the performance of the algorithm, which is an important measure [111]. Vázquez et al. [77] proposed a Fully connected neural network (FCN) for endoluminal scene segmentation. Yu et al. [99] suggested a 3D online and offline CNN for polyp localization. Their approach, however, fails to detect polyps in overexposed and blurred areas in the colonoscopy frame. Depth information is incorporated with the transfer learning frameworks in [112]. Their approach fails to locate polyps that are flat or obstructed. Nguyen et al. [113] proposed a deep encoder-decoder network for polyp segmentation. On the same dataset, they trained and tested their approach. Guo et al. [114] developed a deep hybrid model for polyp identification, namely Dilated ResFCN and SE-Unet. This approach, however, necessitates a high-end computation. Using V-GAN, Pogorelov et al. [115] presented pixel-wise polyp localization. Their approach solves the problem of generalizability in the limited data environment. On the other hand, the article is missing the qualitative performance of the segmentation results. Banik et al. [116] proposed polyp-Net for polyp segmentation. However, the robustness of their method cannot be judged as they used the same dataset for training and testing. Yang et al. [117] proposed mask regions convolutional neural network (MRCNN) for polyp detection and segmentation.

From Table 1.4, it can be seen that different techniques have been proposed for polyp segmentation. The proposed approaches are broadly categorized as (1) handcrafted feature-learning (shape, texture, color, edge, etc.), (2) transfer-learning (pre-trained model), (3) end-to-end learning, and (4) adversarial learning. The handcrafted feature learning approaches do not provide a good polyp segmentation performance similar to the detection problem. The high intra-variations among the polyp features make it challenging to characterize the polyp for segmentation. The challenge of polyp segmentation has been addressed using transfer learning techniques. The learned features of the established models are incorporated in the transfer learning approach, and they are fine-tuned for polyp segmentation. However, because the real-time patient and device-specific colonoscopy images are not suited for transfer learning, the method is not generally suitable for colonoscopy polyp segmentation. As a result, while any model can perform admirably during training, it may fail miserably when it comes to real-time polyp segmentation. The established models like VGG, AlexNet, ResNet, etc., are trained with outdoor natural images. Therefore, the already learned features are sometimes become less valuable while processing the colonoscopy images for segmentation. Therefore, deep models having end-to-end learning supervised learning have been proposed for better feature extraction using varieties

Table 1.4: Overview of existing polyp segmentation approaches in colonoscopy. Abbreviations: LSTM —Long shot-term memory, V-GAN —Variational generative adversarial network

Method	Approach/Technique	Features/Architecture	Datasets
[56]	ellipse fitting techniques based on image curvature, edge distance, and intensity values	Shape	colonoscopy video (Private)
[118]	Multi-scale filtering for noise reduction, suppression of small blood vessels, and enhancement of major edges	Shape	Private
[119]	Edge detection using active contour	Edge detection	Private
[31]	Color wavelet	Texture	Private
[120]	End-to-end semantic segmentation	Enhanced FCN-8, Enhanced variant of SegNet (ESegNet)	CVC-EndoSceneStill
[121]	Pixel-wise initial polyp region candidates prediction by FCN followed by polyp boundaries modeled by texton	FCN-8s	CVC-ColonDB
[76]	Convolution and Transconvolution for feature extractionn followed by reconstruction of prediction map	FCN, U-Net	CVC-ClinicDB
[122]	Information of polyp location and feature	DeepLabv3+LSTM	CVC-ClinicDB
[112]	Transfer learning and End-to-end learning of the existing deep models	FCN-VGG, FCN-GoogLeNet, FCN-AlexNet	CVC-ClinicDB, ASU-Mayo, ETIS-LARIB
[103]	Adversarial learning	V-GAN	CVC-ClinicDB, Kvasir
[71]	Encoder-decoder that uses residual block with squeeze and excitation network	ColonSegNet	Kvasir-SEG
[116]	Dual-tree wavelet pooled CNN (DT-WpCNN)	Polyp-Net	CVC-ColonDB, CVC-ClinicDB
[123]	multi-scale patch-based CNN	CNN	CVC-Clinic DB

of polyp structures. However, getting massive labeled datasets for training such models is very difficult in colonoscopy imaging. Therefore, other learning methodologies like semi-supervised and adversarial learning have been recently proposed. In all, there is a vast scope to explore fairly efficient and robust methods for polyp segmentation.

After polyp segmentation, the doctors finally classify them based on the features they exhibit. They look for geometry features, texture, and color information for detecting cancer in them. Therefore, these image features are extensively used in the proposed automated polyp classifier systems available in the literature.

1.5.3 Polyp classification in colonoscopy images

Handcrafted feature learning-based methods for classification are inevitable when the dataset is in paucity. In [124], Stehle et al. used vascularization features for colon polyp classification in endoscopic images. Condessa et al. [125] used curvature features for detection and classification of colorectal polyps. Fu et al. [126] used texture features taken from both the spatial domain and spectral domain. They applied principal component transform (PCT) and extracted the texture features from the first component of the PCT. A reduced feature dimension set was prepared using sequential forward selection (SFS) and sequential floating forward selection (SFFS) algorithms for classification using SVM. Hafner et al. [127] proposed local texture properties using a 1D histogram using the similarity of neighboring pixels. The compact color vector features were classified using the k-nearest neighbor classifier. Mesejo et al. [128] used a combination of features for a better representation of polyp features. They used texture features and 3D features for the classification of polyps. In [129], Wimmer et al. used wavelet-based features for polyp classification. Engelhardt et al. [130] proposed Color-GLCM features and an SVM classifier for the task. In [131], Bag of words (BOW) descriptors with spatial pyramid matching (SPM) were used. Some deep-learning-based methods and their performances are elaborately discussed in the contributing chapter. These methods discussed above mainly used global features for polyp classification. In most of these methods, a particular feature descriptor is used, which may not be sufficient to characterize polyp features. Also, Most of the existing techniques are validated on a small dataset. The performances of these methods are not satisfactory and cannot be generalized.

Deep learning-based methods can handle such variations better and provide a great generalization. Therefore, there has been a growing interest in using such models in medical image and video pro-

cessing. Byrne et al. [132] used CNNs for the real-time classification of polyps in colonoscopy videos. Ribeiro et al. [133] used image patches to train the CNNs for this task. Later, they fused the hand-crafted features, deep CNNs, and transfer learning features for the task [134]. However, this method would increase the computational burden, and classification accuracy of 93.22% was achieved. The model is also tested on a dataset derived from a high-definition modality, in which numerous image properties such as contrast and tone of vascular tissues are distinguishable, which may not be present in frames captured by standard endoscopic modalities. In [135], to address an imbalanced class distribution, per-class data augmentation is used to improve classification and, as a result, classification accuracy. Golhar et al. [136] proposed a semi-supervised learning system for endoscopic image lesion classification. They incorporated a jigsaw puzzle solver into a semi-supervised learning model that generates discriminative features using an encoder. Their approach has a classification accuracy of 79.76 % when evaluated on their own dataset. We earlier adopted a synthetic image augmentation approach using GAN, followed by a conventional CNN for polyp categorization [137]. The classification accuracy of 88.33% cannot be adequate in this task. This also shows that classical GAN cannot generate authentic colonoscopy-like images. The challenge of polyp categorization has been addressed using transfer learning techniques. The learned features of the established models are incorporated in the transfer learning approach, and they are then finetuned for polyp classification. However, the method is not generally suitable for endoscopic image categorization because the real-time patient and device-specific endoscopic images are not suited for transfer learning. As a result, while any model can perform admirably during training, it may fail when it comes to real-time polyp categorization. Some of the works using the transfer learning approach [138–140], though were deployed, but the desired performances have not been achieved.

Other researchers worked on histopathological images for colorectal polyp classification. Korbar et al. [141] offer a patch-based framework for classifying different types of colorectal polyps from whole-slide images, which was deployed using a ResNet architecture [142]. Wei et al. [143] created a hierarchical classification approach to match the nature of the classification problem for the deep learning model to infer the overall diagnosis of a whole-slide image. Using a sliding window algorithm, each slide was first split down into several patches, and the deep ResNet-like neural network then classified each patch. Song et al. [144] offer a patch-based fully convolutional technique for adenomas classification and grading, with a significant emphasis on model interpretability. From Table 1.5, it can

Table 1.5: Overview of existing polyp classification approaches in colonoscopy. Abbreviations: VAR—variance, EVAR—energy variance, DCT—discrete cosine transform, SSD—Single Shot MultiBox Detector, CNN—Convolutional neural network

Method	Features	Approach	Datasets	Remarks
[124]	Blood-vessel capillaries (vascularization)	Texture	NBI,(37 A + 19 H), Private.	Bleedings or performation cause misclassification
[126]	cooccurrence matrix, VAR, EVAR, DCT	Textural, spatial and spectral domains	No dataset detail	Poor accuracy
[127]	local color vector patterns operator (LCVP)	Textural, color	716 images	Applicable to chromoendoscopy
[130]	Color-GLCM	Textural	NBI, Dye, 466 diminutive polyp images	methylene blue staining and magnification of vascular tissues.
[145]	Blood vessel	Texture	NBI, 434	Not real-time, optical magnification used may not be avilable always
[129]	Wavelet based	Texture	11 different endoscopic databases	Covariance of features extracted from subbands of different color channels seems to be inadequate for the classification of colonic polyps.
[131]	curvature-based feature descriptors , BoW	Texture	CT Colonography images	No quantitative results
[128]	Invariant Gabor texture, Rotational invariant LBP, MPEG-7, Shape-DNA, Kenel-PCA	Texture, color, shape (3-D)	76 colonoscopy videos (NBI,WL)	framework specific and not independent of aparticular technology. Also, validated on small dataset.
[138]	Deep CNNs	Transfer learning	1476 A + 681 H, Private	Low accuracies
[140]	AlexNet	Transfer learning	500 polyp images, Private	Low performances
[132]	Deep CNN	End-to-end training	158 consecutive polyps, Private	Tested on consecutive frames having diminutive polyps.
[133]	CNN	Color, shape and texture	100 images	Lack experimental results comparision
[135]	ResNet50	Per-class data augmentation and CNN	172 A and 31 H	Limited experiments to validate the results
[136]	ResNet-18 and Jigsaw puzzle solver with Siamese network	Semi-supervised	Private	Low accuracies

be seen that a small number of works have been done in the domain of polyp classification. Majority of the works are based on handcrafted deature learning approaches. The non-availability of labeled data limits the exploration of deep learning techniques.

Apart from classifying polyps from 2-D image features, other colonoscopy frames features can be utilized to understand the disease better. The 3-D features of the polyp region can give better polyp surface characteristics. AR and VR applications in this domain can assist the endoscopist in assessing the lesions better. A 3D volume of a polyp in VR is presumably more easily to classify using the Paris classification instead of a 2D image [146]. It can help in accurate resection of the polyp from the normal tissues as it helps in surgical navigation [147]. It provides surgeons with additional anatomical and positional information [148]. The 3-D features are also important in polyp classification [128].

An endoscopic modality generally captures thousands of frames. In this scenario, it is essential to discard low-quality and clinically irrelevant frames of an endoscopic video while the most informative frames should be retained. Therefore, a key-frame selection strategy becomes so important as it saves a lot of reviewing time. Therefore, some works related to automatic key-frame selection have been adopted before poly detection or segmentation. It could assist medical experts and reduce the burden on the clinicians, and helps in better diagnosis. The key-frames have better image features and are thus helpful in better prognosis. A recent work focusing on video summarization is proposed by Li et al. [149]. Clustering-based methods for video summarization in colonoscopy videos are presented in [150, 151]. However, clustering-based methods are not suitable in noise environments. Endoscopy frames are generally susceptible to noise. Also, redundant frames are captured during the endoscopy, making clustering methods perform poorly. Saliency maps for finding key-frames of videos were presented in [152,153]. A color histogram comparison-based method was adopted by Mendi et al. [154]. They compared the color histogram of successive frames in a video sequence, and key-frames were selected using k-means and PCA whenever a significant change in content was observed. However, this model does not fit into endoscopic videos as most of the frames have similar color information. Recently, dictionary learning-based approaches have been proposed for video summarization [155]. In [156], a gastroscopic video summarization technique based on a dictionary learning approach is proposed.

Considering the importance of these aspects in polyp analysis, we also investigated some preliminary works in this thesis. This thesis proposes key-frame selection using depth information of polyps

and subsequent polyp segmentation. The 3-D view of the polyp can be achieved using the selected key-frames. The 3D view gives shape and size information of a polyp. Depth estimation of endoscopic images is a challenging task as the endoscopic images are monocular.

Attempts have been made to solve it as a per-pixel regression problem, however, supervised learning methods require a lot of training data. It isn't easy to acquire depth data without using stereo cameras or expensive depth sensors, as with endoscopy videos. Thus unsupervised methods are being given more importance. Depth estimation in endoscopic video frames imparts clinical relevance to a physician. 3D reconstruction of the monocular images helps in diagnosis and surgical planning.

With this, we try to solve some of the issues related to developing an automated polyp analysis system. The thesis broadly discusses methods for polyp detection, segmentation, and classification. Limitations of the existing techniques and approaches to tackle these problems are discussed. Before going into the objectives and outline of the thesis, the details about the databases available for the studies and for validating the proposed algorithms are worth mentioning. Datasets provide the basis of the methods available in the literature and guide to devise new frameworks best suited. Therefore, the following section discusses the available colonoscopy datasets used for these studies.

1.6 Datasets Description

All the datasets described in Table 1.6 were developed mainly in the late part of the last decade. Also, the datasets made available earlier have limited samples and ground-truth information. Therefore, early existing techniques are primarily based on the hand-crafted feature learning framework. Recently, datasets with a fair amount of labeled image samples have been made available publicly. Therefore, deep learning-based approaches have been newly devised in this domain. CVC-ClinicDB and the Kvsir-SEG are the most used datasets for segmentation found in the literature. The classification dataset [128][3] contains polyp video sequences from three classes, namely, Hyperplastic, serrated, and adenomatous. The UniTOPatho dataset[4] contains polyp histopathological images for colorectal cancer classification and dysplasia grading. The Aichi medical dataset is a private dataset developed for the study of polyp classification[5]. More detailed information about the datasets is provided in the experimental results sections of each approach discussed. Other endoscopic datasets available in this

[3] http://www.depeca.uah.es/colonoscopy_dataset/

[4] http://ieee-dataport.org/open-access/unitopatho

[5] The Aichi Medical University ethical committee has approved this clinical study (January 15, 2018; Approval No. 2017-H304)

Table 1.6: Details of datasets. Abbreviations used: A—Adenomatous, H—Hyperplatic, S—serrated, WL—White light, NBI—Narrow band imaging, VS—Video sequences, H&E —Hematoxylin and Eosin, WS—Whole slide, N—Normal lesion.

Dataset	Organ	Source	Findings	Contents	Purpose
Kvsir-SEG [157]	Large bowel	WL	Polyp	1000 images	Segmentation
SUN Colonoscopy Video Database [158]	Large bowel	WL	Polyps/non-polyps	49,136/109,554	Localization, Detection
ETIS-Larib [78]	Colon	WL	Polyp	196 images	Segmentation
CVC-ClinicDB [159]	Colon	WL	Polyp	612 images	Segmentation
CVC-ColonDB [61]	Colon	WL	polyp	300	Segmentation
CVC-VideoclinicDB [160]	Colon	WL	Polyps/non-polyps	36 VS	Segmentation, Detection
76 Colonoscopy Videos [128]	Colon	WL, NBI	Polyp	A-40, H-21, S-15 VS and frames	Classification
Aichi Medical University (Private)	Colon	WL, NBI, and Dye	Polyps	H-373, A-208 images	Classification
UniTOPatho [26]	Colon	H& E-stained	Polyp Histopathology	292-WS, 41-H, 21-N, 230-A images and 9536 patches	Classification

domain are not useful and are beyond the scope of the current works presented in this thesis.

1.7 Motivation and Objectives

Automated analysis of the colonoscopy images and videos is essential for its usage in a clinical setting. However, colonoscopy frame analysis is a challenging research problem. The task is even more complex and difficult as the images are susceptible to degradation and the presence of high inter- and intra variations of the polyp structures. Most of the state-of-the-art methods are supervised; therefore, they need a lot of labeled samples with variances among the image samples. In addition, many clinical parameter extraction methods focus on estimating the polyps' prominent features. They mostly use important polyp cues like texture, shape, color, etc., features for polyp analysis. However, these global features are sometimes intricate to characterize the polyp features accurately. The local features are vital as they show critical clinical information manifested by the polyp surface. Moreover, the effects of occlusion, geometric transformations, and the impact of noise needed to be analyzed as they may influence the colonoscopy images' characteristics. The constraints of not having large annotated datasets for the training of models must be leveraged by the proposed methods without much performance degradation. Also, approaches best suited for practical applications of polyp analysis are encouraged. Therefore considering these issues and motivated by the above study points, the

objectives of the thesis are given as follows:

- To devise stand-alone methods for automated detection, localization, segmentation, and classification of colonoscopy polyps. To develop integrated frameworks capable of doing one or more tasks simultaneously.
- To detect and localize the polyps in the frames of colonoscopy videos using critical clinical features, temporal information, and geometric features. To provide a method that can perform well for test datasets having different polyp shapes, sizes, textures, and complex backgrounds. To devise systems that can work on stored images and real-time data.
- To devise a method that can be a pre-processing step to polyp detection/segmentation. It will give additional advantages in saving clinicians reviewing time and helps in better diagnosis.
- To propose unsupervised polyp segmentation frameworks that can work on any number of polyp images in case of labeled data scarcity, commonly seen in colonoscopy.
- To devise reliable and data-efficient classifiers for CRC detection from polyp frames and histopathological images. To develop efficient methods that have good generalizability and robustness abilities.
- To integrate all the individual tasks of automated detection, localization, and classification framework and devise an efficient polyp diagnostic assistance system.

1.8 Thesis Outline

To address the issues mentioned in the previous section, this thesis work is organized into six chapters. The contribution of the present thesis is illustrated in Figure 1.7. It includes two major blocks: image acquisition and analysis of the colonoscopy images/videos. Most of the data used in our studies are publicly available. We developed one dataset for polyp classification; the contributing chapter gives the details. The second block represents a flow chart showing different stand-alone and integrated methods for CRC detection. Finally, an approach for further grading of dysplasia in polyps using histopathological polyp images is proposed. The content of each chapter is summarized as follows:

- **Chapter 1** discusses the need for automated polyp analysis systems in colonoscopy videos. Existing works in this view are discussed, highlighting the advantages and limitations of these

methods. The review is broadly divided into three sections: polyp detection and localization, polyp segmentation, and polyp classification. A summary of the proposed techniques and the scope of the thesis work is highlighted.

- In **Chapter 2**, the polyp detection and localization problem are tackled. The first approach uses the dominant polyp features in a modified particle filtering framework to localize polyps. Texture and color information is used in an unsupervised learning approach. Occlusion and specularity, which are commonly seen in colonoscopy, are handled effectively by our algorithm. Finally, active contour (AC) is employed to segment the localized polyps. However, this method gives poor performance for flat and serrated polyps. To circumvent this, our second approach uses both spatial and contextual information—the attention in a YOLOv4 architecture for better polyp detection and localization. Extensive experimental results demonstrate the robustness and generalization of the proposed method.

- In **Chapter 3**, methods for polyp segmentation are devised. The first method is based on the segmentation of prominent and bulged-out polyps, which often show malignancy. This work also discusses various applications of polyp segmentation for better diagnosis. We present an unsupervised polyp segmentation approach using local and global features in the following method. The spatial and contextual information is encapsulated in a proposed adaptive Markov random field (MRF) framework. A shape prior is imposed on the saliency map to generate segmentation masks for polyps in the subsequent work.

- In **Chapter 4**, methods are proposed to classify polyps broadly into benign (hyperplastic) and malignant (adenoma) categories. Local shape and texture features are extracted using efficient descriptors for classification. A fuzzy entropy-based feature selection approach is employed to retain the dominant features. Further, shape features and embedded features extracted using siamese architecture using a triplet network are combined to characterize the polyp features better. Finally, tissue level (histopathological) polyp images are further investigated for cancer grading. In this view, a semi-supervised approach based on a generative adversarial network (GAN) is proposed. This approach is suited to a limited data environment and provides a good classification performance.

- **Chapter 5** summarizes the work presented in this thesis, highlights the major contributions of

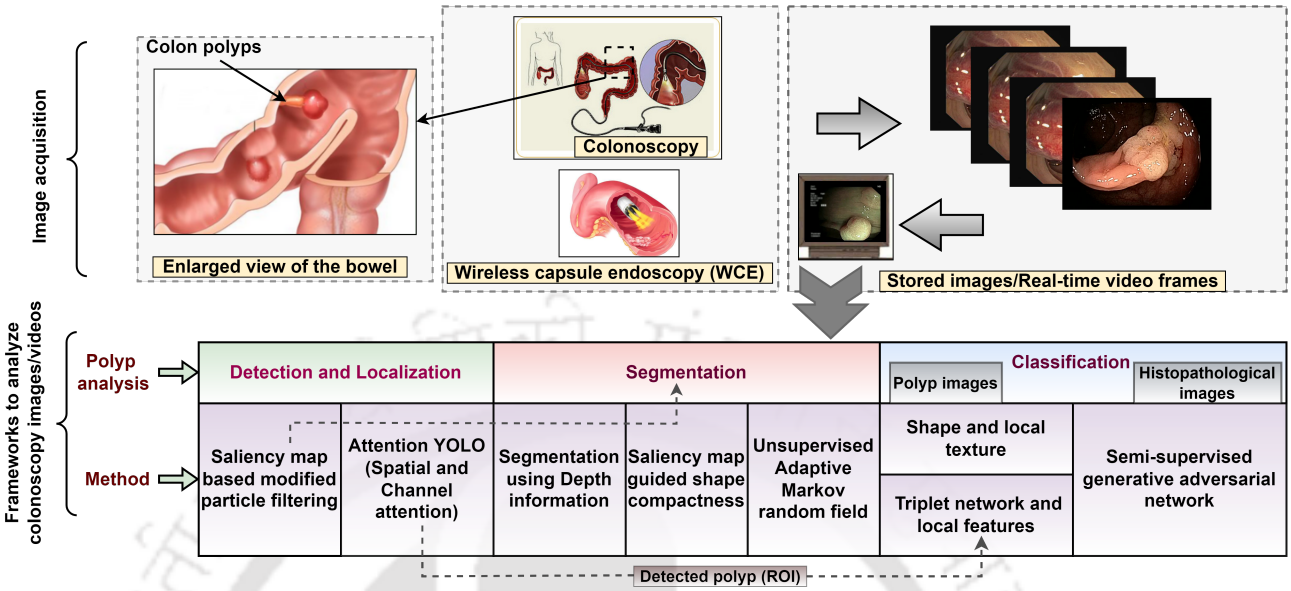


Figure 1.7: Graphical abstract describing dissertation work-flow.

the work, and gives some directions for future research.

2

Polyp Detection

Contents

Objective

This chapter proposes methods for developing automated systems for the detection and localization of polyps in colonoscopy videos. Detection of polyp frames and subsequent polyp localization on video frames are essential steps in diagnosis. This chapter presents techniques for localizing polyps in colonoscopy videos with only polyp frames and polyp and non-polyp frames. Our first work is attempted to localize the polyps in videos containing polyp frames only. It involves the generation of a saliency map using dominant polyp features in an occlusion robust modified particle filter framework to track and localize the polyps in the frames. Further, delineation of the localized polyps is done using the active contour (AC). However, real-time analysis requires processing both polyp and non-polyp frames captured during the colonoscopy procedure. Therefore, the detection of polyp frames is a pretext before the localization of the polyps. For this, attention-based YOLOv4 architecture is proposed for simultaneous detection and localization of polyps. The proposed network captures spatial and contextual information in a better way by giving more attention to the ROIs, which are the clinically significant regions. This approach leverages the polyp miss rate for textureless and minuscule polyps, as seen by the first method. The localized polyps can subsequently be utilized to extract useful features for their classification.

2.1 Introduction

The state-of-the-art polyp detection techniques provide good performance when the colonoscopy images are of high quality with easily distinguishable polyps. However, colonoscopy images are susceptible to geometric and photometric transformations during the colonoscopy procedure. Also, the colonoscope camera's uncontrolled movement may lead to occluded and noisy polyp frames. Another problem that arises during lesion detection in colonoscopy frames is the high variability in the polyp characteristics. Typically, small or serrated polyps, diminutive and isochromatic flat polyps, are missed by the automatic detector. Also, device and patient-specific colonoscopic frames will have different image characteristics [19]. Therefore, the generalization of a particular methodology in colonoscopy image analysis cannot be made. Therefore, all the above-discussed challenges must be considered while proposing an automated polyp detection system.

Both handcrafted feature learning and deep learning-based methods have been proposed over the years for polyp detection in the literature. Handcrafted-based techniques use different cues from the

polyps' image, viz. color, texture, shape, surface properties, etc. On the contrary, deep learning-based methods use the hidden features of the image. Most of the works using handcrafted based feature learning methods are based on supervised learning [61, 161–163]. However, the number of labeled images in this domain is limited. Therefore, these methods provide inferior performances as features learned during the training of a supervised model may not be sufficient to generalize the test images.

All the methods discussed above are supervised. These methods sometimes may not be deployed for localization of polyps due to the requirement of big annotated training images, which is generally difficult in endoscopic procedures. Transfer learning approaches are sometimes adopted in a limited data scenario. During the training phase, the models are trained with large labeled datasets that are substantially different from the test images. Therefore, fine-tuning of these models may not provide desired results on the test images.

Considering all the challenges discussed above, we propose methods that can be suitable in different scenarios with acceptable performances. Our proposed polyp detection systems can be employed for offline (stored) video data. Polyps contain the clinical information for cancer diagnosis. Therefore, only polyp frames are retained, and uninformative frames (non-polyp frames) and discarded while acquiring images. Subsequently, the clinicians analyze the polyps (ROIs) to detect abnormalities. However, polyps and non-polyp frames are acquired during the colonoscopy procedure in some scenarios, e.g., WCE captures around 50,000 images. Sometimes the doctors let the camera systems capture both the varieties of frames and subsequently post-process to detect and localize polyps in them.

Considering all the scenarios, methods for finding polyps in colonoscopy videos with only polyp frames and having both polyp and non-polyp frames are proposed in this chapter. The proposed saliency map-based tracking framework for polyp detection is proposed in section 2.2. Section 2.3 describes our proposed approach based on attention YOLOv4 model. Finally, a summary is provided in section 2.4.

2.2 Saliency Map-Based Modified Particle Filter

The proposed method uses visual saliency as the backbone for a modified particle filtering framework[1]. It tracks and localizes the polyps on each frame of the colonoscopy video sequence. The proposed visual saliency framework tries to model the human visual system and polyp features. Our method employs

[1]This work has been published in IEEE Transactions on Instrumentation and Measurement, 2021 (Refer *List of publications* page for details).

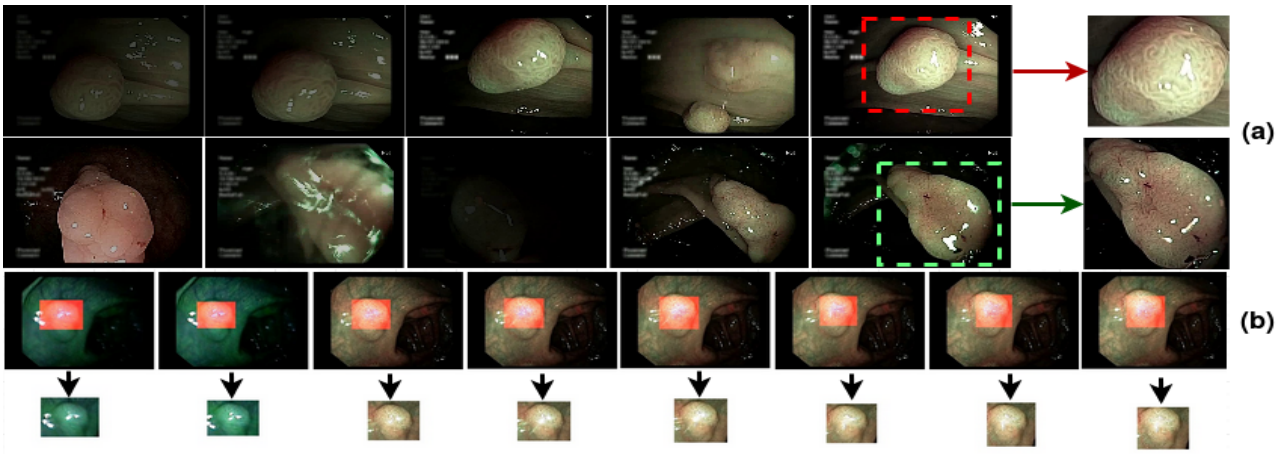


Figure 2.1: Localization of polyps in colonoscopy video frames. (a) manual ROI selection, (b) automated localization of ROIs.

an unsupervised approach. The shape information of the polyp is also incorporated into an active contour framework for final localization and subsequent segmentation. During the colonoscopy, the frames sometimes suffer from occlusion and are out of focus for a few consecutive frames. Such polyps hardly show any clinical manifestation and thus require no attention. A novel forward prediction filter is incorporated into our framework to leverage this. This concept is reused in our previously published proceeding [164]. Thus, a detection and localization system would be immensely helpful in managing big medical data by assisting medical experts. It will have many application areas in this domain. Mainly, the selected polyps are analyzed to classify them into benign and cancer types.

The summary of our proposed work is as follows: This work presents a saliency map-based polyp region localization in colonoscopic video frames. Our method uses the saliency map or probability map as a measurement model for the particle filtering-based tracking framework. The shape of the polyps is used by the particles of the tracker for final refinement using an AC model.

2.2.1 Proposed ROI selection framework

Detection and localization of polyp in endoscopic video frames has a lot of medical applications. Generally, this task is done under an expert's intervention. As shown in Figure 2.1(a), the polyp is localized by putting a bounding box around the ROI. Nowadays, a computer-aided diagnostic (CAD) system is used by physicians to automate this process. The real challenges incurred for automatic detection come from high variability in the frames, as can be seen in Figure 2.1. Variations of polyp shape, size, geometry, and other inherent characteristics put a lot of challenges in CAD. Nevertheless,

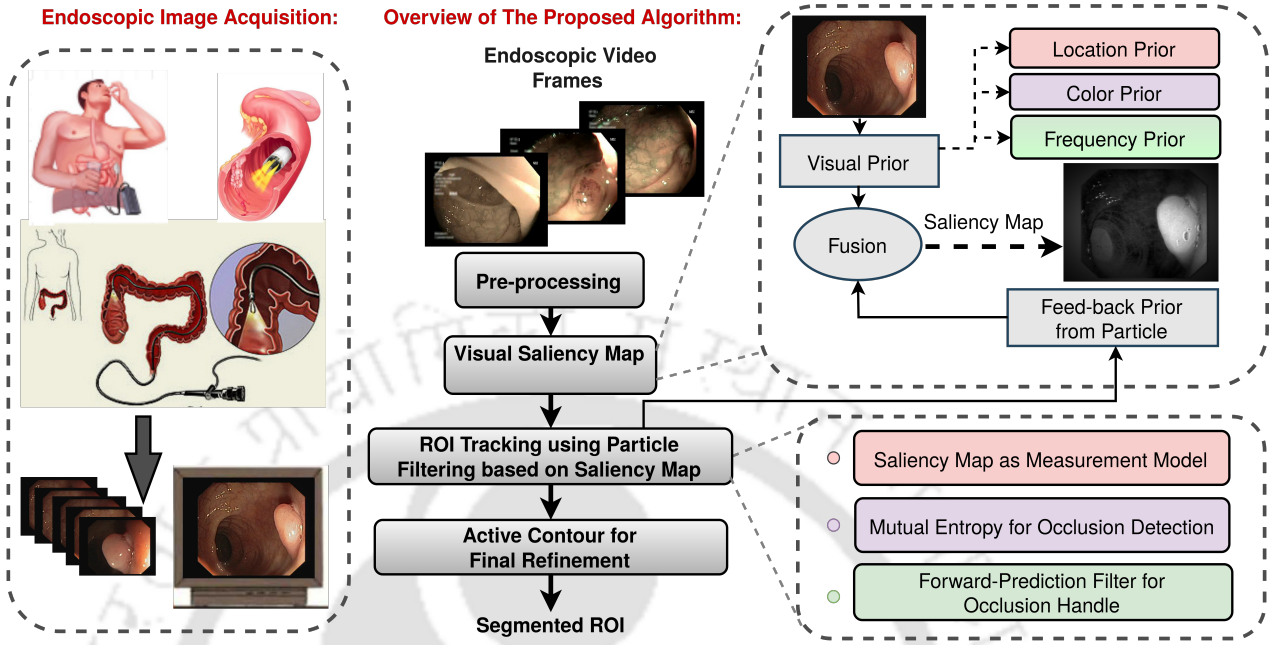


Figure 2.2: Overview of the proposed method for localization of polyps in the colonoscopy videos.

motion blur, high specularity, occlusion, and out-of-focus are very common during colonoscopy, making localization difficult. In our work, the localization of ROI is considered as an object tracking problem, as shown in Figure 2.1(b). Our method adopts a novel occlusion robust tracking framework for polyp localization on the back-bone of the saliency map. The overview of the proposed method is shown in Figure 2.2.

Different object tracking frameworks have been proposed over the years, viz tracking learning detection (TLD) [165], correlation-based tracking [166], particle swarm optimization (PSO) [167], etc. Particle filter in object tracking has been used extensively recently. Some of the tracking approaches based on particle filtering are proposed in [168–170]. All the methods discussed have used different measurement models of the particle filter. On the contrary, our proposed method proposes using a robust modified particle filtering framework to handle occlusion for any measurement model. An uncertainty factor is introduced to control the search space of the filter particles. Whenever the particles lose the target, they gradually increase their search area until they find the target again.

2.2.1.1 Modified particle filter framework

During the colonoscopy, the captured images may suffer from blurring and occlusion. In such a scenario, tracking an object becomes very difficult. Detection of such events during colonoscopy is

very crucial in polyp detection. This work introduces an uncertainty factor to detect occlusions or blurring. Cumulative measure by all the particles in the particle filter is taken for each frame of the colonoscopic video sequence. In this work, mutual entropy as a measure is proposed. Mutual particle entropy, H is given by:

$$H = \sum_{i=1}^{N} w_k^i log(w_k^i) \tag{2.1}$$

where, w_k is the weight of particle at kth instant. When the particles converge to the target, mutual entropy tends to decrease, and it increases when the particles lose the target either because of occlusion or blurring. Thus, an uncertainty factor is introduced, which is related to mutual entropy given by:

$$C = 1 - e^{-H^\gamma} \tag{2.2}$$

where, γ is a constant lying between 0 and 1.

2.2.1.2 Particle update

For a particle filter framework the state update equation is given by:

$$\vec{X_k} = \vec{X}_{k-1} + k\vec{\mathcal{N}} \tag{2.3}$$

where, $\vec{\mathcal{N}}$ represents the measurement of noise and k is a scaling constant. In this work, the modified framework has the state update equation given by:

$$\vec{X_k} = \vec{X}_{k-1} + C\vec{\mathcal{N}} \tag{2.4}$$

where, $C = 1 - e^{-H^\gamma}$. Thus, for a very confident measurement, the noise gets scaled down to a low value, and the search space becomes small and vice versa.

2.2.1.3 Measurement

The weight assigned to each particle depends on the closeness to the neighborhood target features. If the measurement is closer to the target, then higher weights will be given to the particles. These weights give the probability of the presence of the target in the neighborhood of each particle. Different measurement models have been proposed in the literature. H. Yao et al. [171] proposed using a combination of the color moment model and texture model for measurement in face tracking applications. A saliency map generated for each frame is used for measurement in our work. The weights assigned to the particles in the neighborhood of the salient regions, i.e., the polyp regions, are higher

than the weights of particles associated with the non-salient areas.

2.2.1.4 Occlusion handle

A forward prediction filter is activated when occlusion is detected using the uncertainty measure. It predicts the next state as a combination of a few previous states. The details of the method can be found in our paper [164].

2.2.1.5 Re-Sampling of particles

This is the final step in every iteration of the tracking algorithm using the particle filter approach. Here, N particles are randomly picked from the existing particle set according to the updated weights. The particles with higher weights will be picked more than once, and finally, the target will be localized.

2.2.1.6 Saliency map based ROI extraction

In the case of localization of polyp in endoscopic videos, the use of a visual saliency map as a measurement model is proposed in this work. The concept of visual saliency tries to map the image in a way the human eye would perceive. It also tries to encapsulate the inherent characteristics of a polyp for map generation. Thus, different maps will be generated by quantifying a few of such characteristics of a human eye and the attributes of polyps. These maps, when used together, can highlight the salient portion of an image. Zhang et al. [172] suggested a novel method for saliency detection, namely SDSP, that generates the saliency map by combining simple priors. Our work uses a modified form of this concept to generate a saliency map by including a few more priors, which is explained in brief in the later sections. The saliency map is very sensitive to specular reflections. In this work, we suggested using particle variance from the previous frame as a prior along with the other priors to restrict the specular surfaces around the target to affect the saliency map. This technique makes the localization process unsupervised, i.e., no previous target is required to learn the object being tracked. The concept of visual saliency is extensively used in the field of segmentation. Our experiments show that this is an advantageous measurement model in endoscopic videos.

In the suggested method, the map is generated by combining frequency prior, color prior, and location prior. All computations for the three priors are done in CIEL*a*b color space. The L component signifies the lightness, a* signifies the green-red color component, and b* represents the blue-yellow color component. This work incorporates a feedback system into the SDSP technique to generate a new map that can penalize the unwanted areas highlighted by the SDSP technique.

Our experiments show that the SDSP technique is susceptible to specular reflections. In endoscopic videos, the specular surfaces affect the tracking performance. So, penalizing the unwanted area by using information given by the particles in the previous frame significantly improves the tracking performance.

Frequency prior: By nature, our eye has a bandpass filter. Thus, only a specific range of color frequency can be allowed to pass. A log Gabor filter models the bandpass filter. The choice of this filter is because the log Gabor filter has an extended tail at the high-frequency end, making it closer to the response on an eye. In the frequency domain, the transfer function of this filter is given by:

$$G(\mathbf{u}) = \exp\left(-\left(\log\frac{\|\mathbf{u}\|_2}{\omega_0}\right)^2 / 2\sigma_F^2\right) \tag{2.5}$$

where, $\mathbf{u} = (u, v)$ represents the coordinates in the frequency domain, ω_0 represents the center frequency of the filter, and σ_F^2 represents the variance about the center frequency. The filtered output in CIE-La*b* color space is denoted by $f_L(\mathbf{x})$, $f_a(\mathbf{x})$, and $f_b(\mathbf{x})$, representing the three channels. Now, to generate final map from the frequency prior, the three channels are combined using the following equation:

$$S_F(\mathbf{x}) = ((f_L)^2 + (f_a)^2 + (f_b)^2)^{\frac{1}{2}}(\mathbf{x}) \tag{2.6}$$

Color prior: Zhang et al. found that some studies suggested that the human eye better perceives warm colors like red and yellow as compared to cold colors like green and blue [172]. Generally, the color of an adenomatous polyp (malignant polyp) is more likely to be deep red or purple, and the color of a non-adenomatous lesion (benign polyp) tends to be yellow or white [16]. This property of a polyp is encapsulated by our proposed color prior. In CIE-La*b* color space, the pixel with a greater a* value would appear reddish, and a greater b* value would appear yellowish. Let $f_a(\mathbf{x})$ and $f_b(\mathbf{x})$ denote the a* and b* channel values at the position, $\mathbf{x}$ respectively. Firstly, these values are linearly mapped into values between 0 and 1, i.e. $(f_{an}(\mathbf{x}), f_{bn}(\mathbf{x})) \in [0, 1]$ by the follwing relation:

$$f_{an}(\mathbf{x}) = \frac{f_a(\mathbf{x}) - min\text{a}}{max\text{a} - min\text{a}}, f_{bn}(\mathbf{x}) = \frac{f_b(\mathbf{x}) - min\text{b}}{max\text{b} - min\text{b}} \tag{2.7}$$

where, maxa, mina, maxb, minb stands for the maximum and minimum values of $f_a(\mathbf{x})$ and $f_b(\mathbf{x})$ respectively. Finally, the target is to enhance the pixel with greater a* and b* values and penalize the

pixels with lower values. This is done using the following relation:

$$S_c(\mathbf{x}) = 1 - \exp\left(-\frac{f_{an}^2(\mathbf{x}) + f_{bn}^2(\mathrm{x})}{\sigma_c^2}\right) \tag{2.8}$$

where, σ_c^2 is a constant.

Location prior The human eye tends to get attracted to objects near the center of the image frame. Thus, a map needs to be generated, which enhances the pixels in accordance with their distance from the center of the frame. This can effectively be modeled as a Gaussian with its mean at $\mathbf{c}$. This can be given by:

$$S_D(\mathbf{x}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{c}\|_2^2}{\sigma_D^2}\right) \tag{2.9}$$

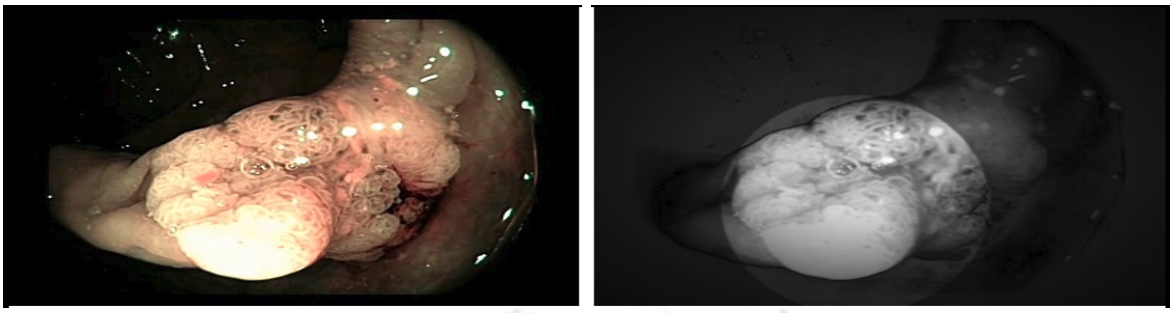
where, σ_D^2 is the variance of the Gaussian and can be treated as a parameter.

Individually these priors generate different maps. The output maps generated for each prior and with the combined priors are shown in Figure 2.3. The final map is generated by pixel by pixel multiplication of each map.

2.2.1.7 Feedback from particles

In this section, we suggest a way to suppress the unwanted portion that the SDSP highlights. As mentioned earlier, the SDSP method is susceptible to specular reflections, which are very commonly found in endoscopic frames. In the localization process, the particles tend to decrease their variance as they converge to the target. Thus, as soon as the convergence occurs, the target will be enclosed in an elliptical body whose major and minor axis will be given by the variance of x and y coordinates of the particles, i.e., σ_x^2 and σ_y^2 centered at the centroid of the particles. The assumption of the target's elliptical shape is based on the fact that polyps have elliptical shapes or curved edges. Thus, in each frame, an ellipse with the variance of the particles can be generated. The pixels lying outside the ellipse will be penalized and carry a smaller weight than those lying inside. Intuitively, it can be said that in the first frame when the particles are all over the image, almost all the pixels lie in the ellipse. As iterations proceed and the particles localize, only the target pixels fall inside the ellipse, and the unwanted regions highlighted by the SDSP technique get suppressed. Thus a pixel with coordinates (x, y) will lie inside the ellipse if:

$$\frac{(x - c_x)^2}{\sigma_x^2} + \frac{(y - c_y)^2}{\sigma_y^2} < 1 \tag{2.10}$$



(a) (b)

Figure 2.3: Final saliency map generation including the feedback structure from the original image; (a) original image (b) saliency map.

where, (c_x, c_y) is the centroid of all the particles. We call the map generated from the feedback structure as $S_S(\mathbf{x})$.

Now, the final saliency map is given by :

$$V(\mathbf{x}) = S_F(\mathbf{x}).S_c(\mathbf{x}).S_D(\mathbf{x}).S_S(\mathbf{x}) \tag{2.11}$$

An example of the whole process is shown in Figure 2.3. It can be seen that the region containing the polyp is highlighted in the saliency map.

2.2.1.8 Measurement model

Any method that can learn some target features and classify between target and non-target can be a measurement model. Traditional hand-crafted features do not characterize the polyp across frames. Thus, a probability map-based ROI selection is proposed. As we know that polyp is a visual anomaly, the saliency map will highlight the polyp by giving the pixels in the affected area a higher weight. In the framework, the particles can pick the values returned by the saliency map and treat them as the measurement value. The values can then be normalized to give the final updated particle weights. The variance of the particles in x and y directions will also give the size of the bounding box where the ROI lies.

2.2.1.9 Segmentation using adaptive mask formation for active contour

The deformable model is one of the most promising object detection, segmentation, and localization methods, especially in medical imaging. This method exploits image features like location, shape, size of the object apriori. Such type of model gives better results in medical image segmentation as the

prior knowledge about the shape and size of the object is known apriori. In this work, a region-based model is adopted, which is known as Chan-Vese Model using the level set formulation [173].

The performance of active contour depends quite a lot on the initial mask. Better the initial approximate better will be the convergence. The particle filter framework with a visual saliency map as a measurement model localizes to the required section of the image. The initial shape of the mask is assumed to be elliptical, whose major and the minor axis will be given by the variance of x and y coordinates of the particles, i.e., σ_x^2 and σ_y^2 centered at the centroid of the particles. With subsequent iterations, as the particles converge, so does their variance in both directions. Thus, after a few iterations, we get a close approximation to the original shape, and thus the performance of the active contour improves. Figure 2.4 shows the results of segmentation using active contour at each level of mask generation. The input image is the same as the image given in Figure 2.3. But, as the iteration of the particle filter proceeds, the elliptical mask starts evolving around the polyp structure, giving the best approximate of the target. Thus, much better segmentation results in the final frames are obtained.

The energy function used in AC is iteratively evaluated until the difference between the previous and current segmented area becomes stable. In our experiment, it was seen that there was not much change in convergence after some initial iterations. The number of iterations that can be reliably taken is 12 for the final convergence of the AC, as can be demonstrated with an example shown in Figure 2.4.

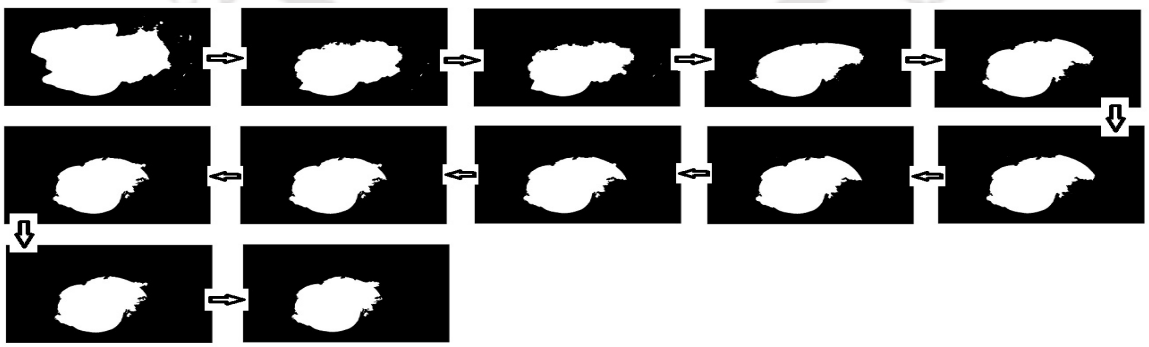


Figure 2.4: Results of segmentation in 12 iterations of adaptive mask formation using particle filter.

2.2.2 Results and discussion

For the experimental work, the video sequences used in this work are taken from a publicly available dataset [128] which can be found in the following link: http://www.depeca.uah.es/colonoscopy_dataset/. It contains video sequences from three polyp classes; adenoma, hyperplastic, and serrated. The dataset contains NBI and WL images.

For evaluation and comparison of the segmentation performance of the proposed method, the CVC-ClinicDB database [159] is used. This publicly available dataset contains 612 colonoscopic frames from 29 video sequences. The tracking of polyps with the proposed individual priors and combined priors is shown in Figure 2.5. The green and the yellow-colored arrows of Figure 2.5 represent partially tracked polyps when all the priors are not combined to generate a saliency map. Finally, refinement in localization is further achieved by fusing the feedback prior to these priors followed by the AC. Parameters for the generation of saliency maps were empirically tuned. The tracking results and extracted ROI of some of the frames of [128] are shown in Figure 2.6.

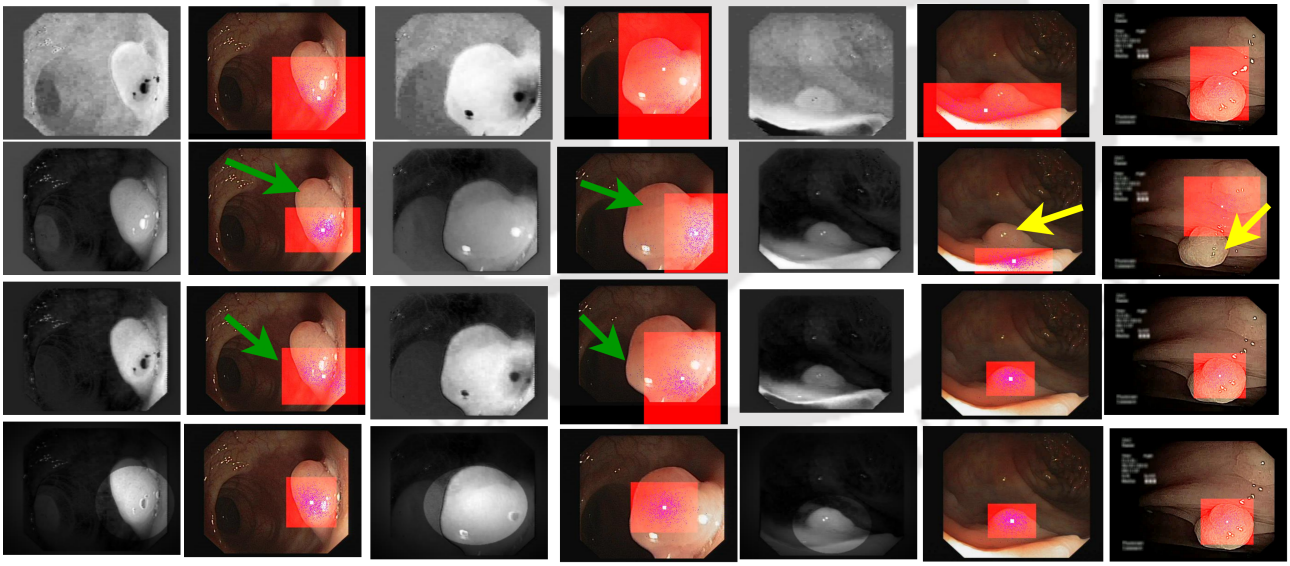


Figure 2.5: Saliency map generation and tracking results using different priors; 1^{st} row: only color prior; 2^{nd} row: only frequency prior; 3^{rd} row: both color and frequency priors; 4^{th} row: combination of all the priors which shows that every prior is important for polyp localization.

Each video sequence consists of several frames captured with 30fps. Our algorithm takes a frame as an input, generates a saliency map, and subsequently localizes poly in the images. To validate the performance of our method for different imaging modalities, we tested our method on this database as it has long video sequences of colonoscopic polyp frames of both NBI and WL imaging. The average

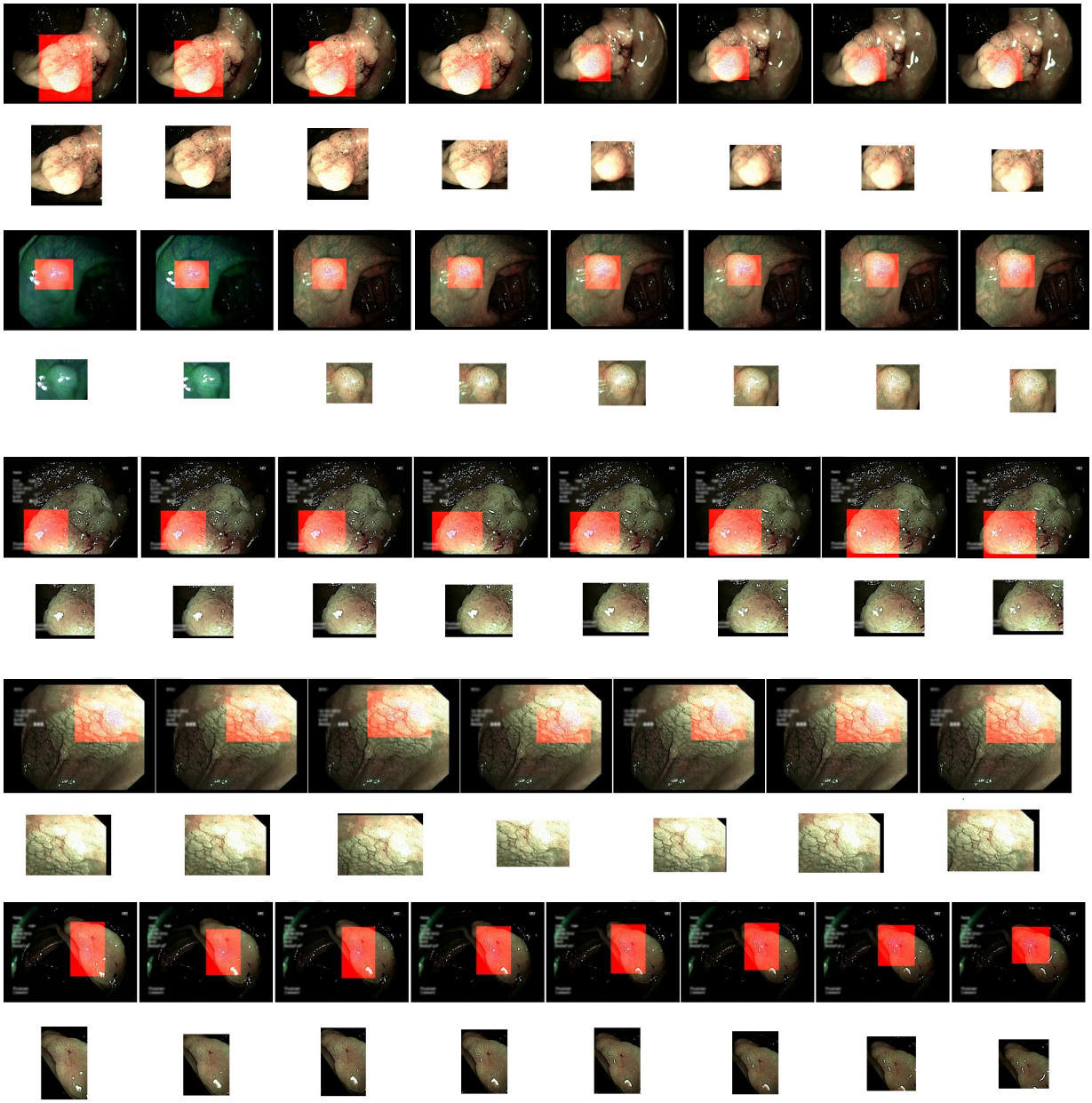


Figure 2.6: Tracking results on a few frame with its corresponding extracted ROI in the second row.

Table 2.1: Tracking efficiency for some video sequences of NBI

–	Video V1	Video V2	Video V3	Video V4
Efficiency	98.3108%	91.9159%	96.3210%	99.2521%

–	Video V5	Video V6	Video V7
Efficiency	92.6910%	49.3023%	97.8902%

Table 2.2: Segmentation score for each video sequences of NBI

–	Video V1	Video V2	Video V3	Video V4
Score	75.74%	81.99%	83.62%	47.83%

–	Video V5	Video V6
Score	70.00%	30.34%

processing time in second(s) for processing a 256×256 image frame by our algorithm is 32.27 sec/frame. The experiments were performed on a system with an Intel i5-4570 CPU @ 3.20GHz×4 and an 8 GB RAM. Our proposed system can provides different outputs like tracking score, ROI, saliency map, and segmentation score for each of the polyp frames.

Our work aims to return the ROIs from each frame so that the output images can be used to classify polyps. For this task, 50% area of the ROI must contain the polyp. In this work, tracking efficiency is used as the measurement for performance. The tracking efficiency is calculated by counting all the ROIs whose 50% is filled with the polyp. Thus, tracking efficiency is the ratio of that number to the total number of frames in the video sequence. Table 2.1 and Table 2.2 show a quantitative measure of the tracking performance and segmentation performance of some of the video sequences of NBI, respectively. Similarly, Table 2.3 and Table 2.4 show these measures on the frames of WL imaging. The consistency in performances for both of the imaging modalities shows the robustness of our method.

Table 2.3: Tracking efficiency for some video sequences of WL

–	Video V1	Video V2	Video V3	Video V4
Efficiency	96.033%	97.556%	97.34545%	94.4566%

–	Video V5	Video V6	Video V7
Efficiency	93.5657%	63.466%	97.466%

From the visual results given, we can infer that specular reflections play a vital role in the local-

Table 2.4: Segmentation score for each video sequences of WL

–	Video V1	Video V2	Video V3	Video V4
Score	72.23%	85.99%	81.62%	45.83%

–	Video V5	Video V6
Score	68.45%	40.34%

ization process. Specular surfaces are generally classified as salient regions. The main motive of using the feedback structure in the saliency map is to remove the wrongly classified regions.

The average segmentation score is generated for each video sequence as quantitative analysis. This is the Jaccard index (J) which can be computed as follows:

$$\text{Dice } (D) = \frac{2 * TP}{2 * TP + FP + FN} \tag{2.12}$$

$$\text{Jaccard index } (J) = \frac{D}{2 - D} \tag{2.13}$$

where, $TP, FP,$ and FN are True positive, False positive, and False negative, respectively. These metrics are calculated based on the pixels. A GUI designed using the proposed algorithm for automatic polyp tracking and segmentation in an endoscopic video sequence is shown in Figure 2.7.

our proposed method is the first to detect polyps in endoscopic video frames in a particle filtering framework to the best of our knowledge. Though the segmentation score is less than the current deep model-based methods, our model is advantageous. The deep model-based methods have many limitations. All the deep models are trained with the same datasets and tested with the same datasets. Also, computational complexity is very high for these models. All the models can only be applied to stored endoscopic images. On the contrary, our method is unsupervised and needs no prior learning. It can be applied to online endoscopic video sequences for polyp detection and segmentation. The dataset used for the comparison is the CVC-ClinicDB database [159].

Dice coefficient and Jaccard index are used as the segmentation performance measures. Figure 2.8 shows a qualitative performance of some samples of the CVC-ClinicDB database. Table 2.5 shows the dice score of different baseline models used in the medical image processing domain and other individual models on the CVC-ClinicDB database. From Table 2.6, it can be seen that our proposed work gives a competitive performance with most of the state-of-the-art methods. We also validated our method with another publicly available dataset, which is ETIS-Larib. This dataset contains 196

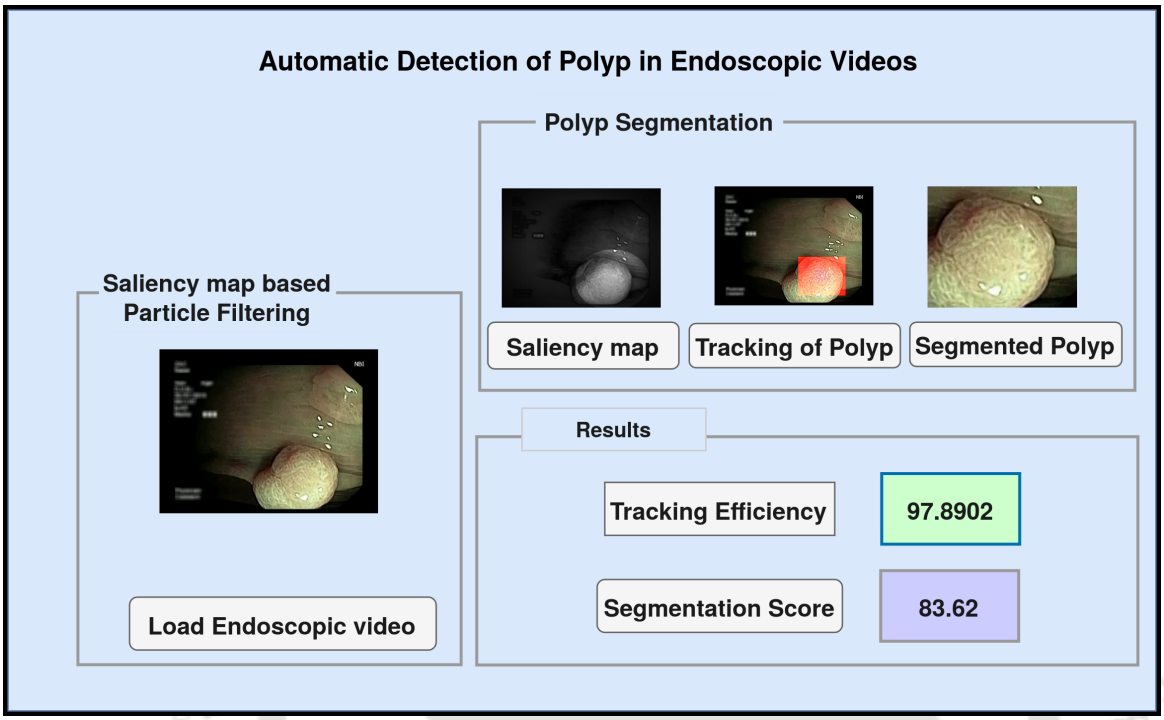


Figure 2.7: Designed GUI for detection and segmentation of Polyps; an example endoscopic frame of a video sequence showing performance of the proposed automatic detection system.

colonoscopic frames. The average mIoU achieved in polyp segmentation using our method is 46.28%.

Our proposed model can be used as a DAS in polyp analysis. After polyp segmentation, an endoscopist analyzes image features like texture, color, geometry, and shape to detect polyps malignancy.

In this direction, we earlier proposed an automatic polyp classification system with the help of features extracted from the segmented polyp [174]. The polyps were first detected and then segmented by Prof. Kunio Kasugai of Japan's Aichi medical university and hospital. Another application of our method could be the selection of keyframes of a long colonoscopic video on the basis of segmented polyps.

Table 2.5: Comparision of segmentation performance with differfent baseline models and other individual models on CVC-ClinicDB database

Method	Dice (D)
DT-WpCNN [116]	0.809
LGWe-LSM [116]	0.754
U-Net [175]	0.767
Resnet50 [142]	0.718
Hybrid-CNN [114]	0.834
Polypnet [116]	0.839

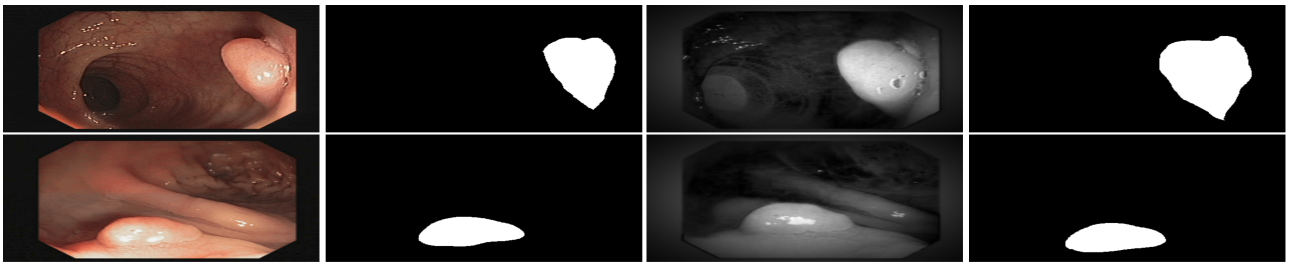


Figure 2.8: Segmentation results on some of the frames of CVC-ClinicDB database. col 1: pre-processed image, col 2: ground-truth mask, col 3: saliency map, and col 4: obtained segmentation mask.

Table 2.6: Comparative analysis with state-of-the-art methods on the CVC-ClinicDB Database

Method	Dice Coefficient (D)	Jaccard index (IOU)
MSA-DOVA [176]	36.27	22.13
SA-DOVA [159]	55.33	37.93
FCN8 (2 class) [77]	67.44	50.85
Shape-UCM [161] (87 polyps)	65.77	49.00
GPB-OWT-UCM [177] (87 polyps)	61.11	44.00
Proposed (612 polyps)	66.06	49.25

2.2.3 Conclusion

This work proposes to use a saliency map in a tracking framework for polyp localization and segmentation in colonoscopy videos. A new measurement model is proposed, which is the visual saliency map. This work also shows how the saliency map can be used to choose particle weights and makes the tracking process faster. The inherent features of a polyp are used for the generation of such maps. The shape of the polyp is used for refinement of the ROI using an active contour model. It thus helps in discarding the specular regions and converges towards better delineation of polyps. The experimental results show that our method is competitive with the state-of-the-art techniques in polyp segmentation. An experimental study on the CVC-ClinicDB database using the proposed method shows that our method can be used effectively in polyp localization. Our approach is used for offline endoscopic video processing. The proposed method sometimes fails to localize small patchy polyps and polyps in highly over-exposed regions. As discussed earlier, a better saliency map can achieve a better localization performance; Therefore, in our subsequent work, we propose using the hidden discriminating features from various polyp structures to generate the saliency map for polyp detection. Also, we aim to achieve real-time colonoscopy video processing for clinical applications.

2.3 Attention based YOLOv4 Framework

Because of the generalization property, deep learning-based approaches have been extensively used in medical image analysis, especially colonoscopy image analysis. Many deep-learning-based approaches for polyp analysis using colonoscopy videos are available in the literature, showing promising performance. A detailed analysis of different deep learning approaches for colon cancer is available in the literature [178]. They discussed the methodologies, advantages, and disadvantages of these techniques in five separate studies: polyp detection, classification, segmentation, survival prediction, and inflammatory bowel diseases. However, This work discusses the importance, advantages, and disadvantages of deep-learning frameworks for polyp detection.

Polyp detection systems based on deep learning have improved overall performance in video frames of colonoscopy [72]. One significant advantage of the deep learning-based techniques is that we can establish the generalization ability of these models. In medical procedures, especially in polyp detection and localization systems, the following features are desired: 1) consistency in performance, i.e., the DAS must reliably produce the performance independent of imaging modalities and patients. 2) minimum polyp miss rate, i.e., a high detection rate, and 3) real-time application, which could help early diagnosis.

Considering all these requirements for devising an automated polyp detection system, we propose an attention based YOLOv4 framework. The main contribution of our proposed method can be summarized as follows. This work presents an attention mechanism in the YOLOv4 framework for improved polyp detection. Our approach proposes to use spatial and channel attention modules in the YOLOv4 framework. The attention modules give importance to the region of interests (ROIs), i.e., the polyp regions in the colonoscopy frames. A comparison of performance based on important matrices with state-of-the-art methods is presented in this work. The performances evaluated on two databases validates the robustness and generalization capability of our approach. The detailed analysis of this approach is presented in section 2.3.1. The rest of the work is organized as follows: Section 2.3.1 presents the proposed methodology. Experimental results are given in section 2.3.4. Finally, section 2.3.5 concludes the proposed work.

2.3.1 Proposed method

Before proceeding to the methodology, details of the datasets are essential. As discussed earlier, training the model with varieties of polyp structures is crucial to reducing the detector's polyp miss rate. Also, it assures achieving the generalization and robustness capability of the proposed model.

2.3.2 Dataset

We used two databases, viz. 1) Kvsir-SEG [157] and 2) SUN Colonoscopy Video Database [158] for polyp detection. The Kvsir-SEG database is a freely available open-access database. SUN (Showa University and Nagoya University) Colonoscopy Video Database is the dataset designed to evaluate an automated colorectal-polyp detection system. This database is also publicly available. It comprises 49,136 polyp frames taken from 100 different polyps using a high-definition endoscope (CF-HQ290ZI and CF-H290ECI; Olympus, Tokyo, Japan). Similarly, the Kvsir-SEG dataset contains 1000 image frames acquired using ScopeGuide, Olympus Europe, endoscope. Some of the samples from both the datasets are shown in Figure 2.9. The details of the datasets are given in Table 2.7.

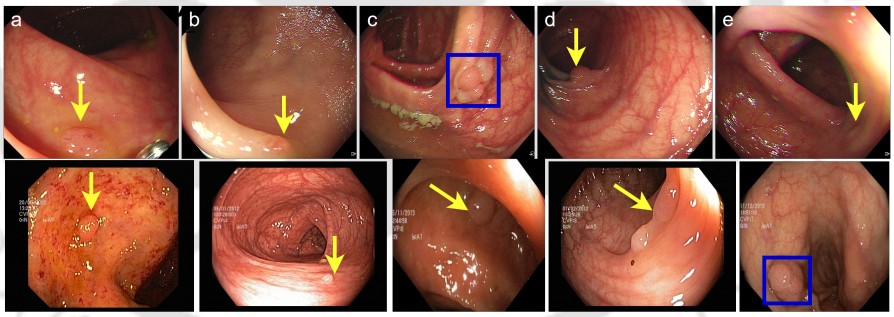


Figure 2.9: Some of the representative images from the trained databases. First-row image samples are from the SUN Colonoscopy Video Database, and second-row images are from the Kvasir-SEG dataset. (a) 18 mm high-grade adenoma. (b) 2mm hyperplasic diminutive polyp. (c) 10mm low-grade adenoma polypoid polyp. (d) 4 mm distant diminutive polyp. (e) flat polyp.

Table 2.7: Details of the datasets.

Dataset	Organ	Findings	Dataset Contents	Size/ Polyp morphology
Kvsir-SEG [157]	Large bowel	Polyp	1000 images	Large polyp: 700 ($>$ 160 $\times$ 160 pixels) Median polyp: 323 ($>$ 64 $\times$ 64 pixels and $\leq$ 160 $\times$ 160 pixels) Small polyp: 48 ($\leq$ 64 $\times$ 64 pixels)
SUN Colonoscopy [158]	Large bowel	Polyps/non-polyps	49,136/109,554	Median (IQR) mm: 5 (3-7) Diminutive polyp ($<$ 5mm): 60 Morphology (Protruded/ flat): 66/ 34

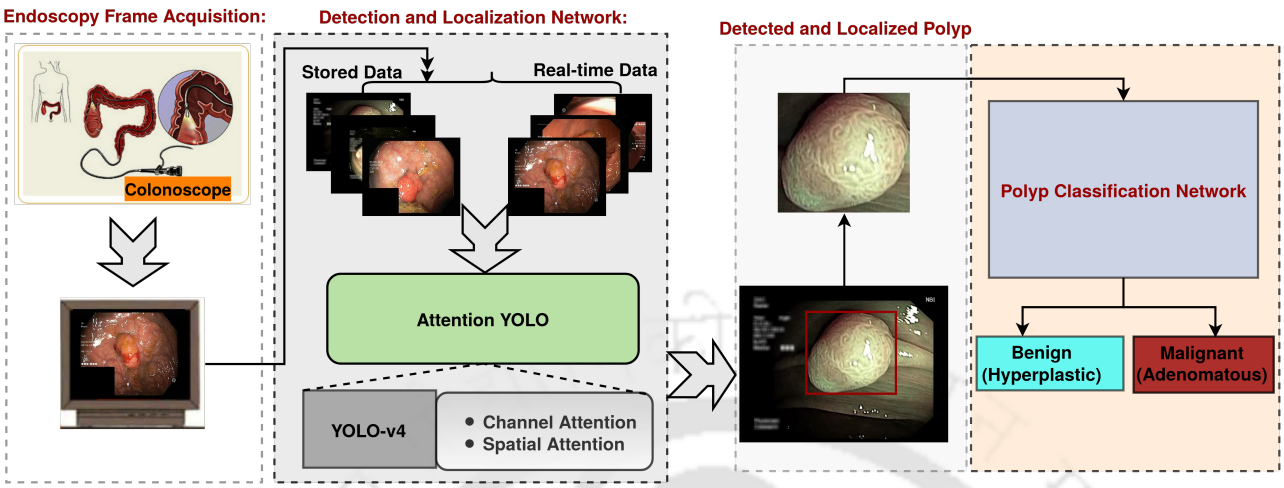


Figure 2.10: Proposed algorithm.

2.3.3 Proposed polyp detection method

Figure 2.10 shows the overall schema of our proposed methodology. The last block, i.e., the polyp classification network, is used to classify the detected polyps. This part of the work is presented in the next chapter. During the colonoscopy, the captured video frames are stored in a computer system for further analysis in the future. However, real-time analysis of colonoscopy video frames can lead to better diagnosis and early treatment. Also, the decision support system must automatically detect any abnormality in the frames. Therefore, the first stage of our current work focuses on handling real-time data for the automatic detection of polyps in the colonoscopy frames. A deep learning-based attention YOLOv4 model is proposed in this work[2]. The architecture of the proposed model is shown in Figure 2.11. Furthermore, the localized polyps are classified as adenoma or non-adenoma employing our suggested classification network. The classification approach is explained in further detail later in this thesis.

2.3.3.1 Attention YOLO

YOLO is a single-stage object detection model and is considered superior to other deep learning models owing to its optimal accuracy and detection speed [83]. Further, YOLOv2 [179] and YOLOv3 [180] were proposed, which show improved detection performances. In YOLOv3, a CNN-based-Darknet53 is employed as a backbone of the architecture, efficiently extracting features from the input image. Later, YOLOv4 was proposed by Bochkovskiy et al. [181] to enhance the detection performance and

[2]A revision for this work has been submitted to Scientific Reports, Nature (Refer *List of publications* page for details).

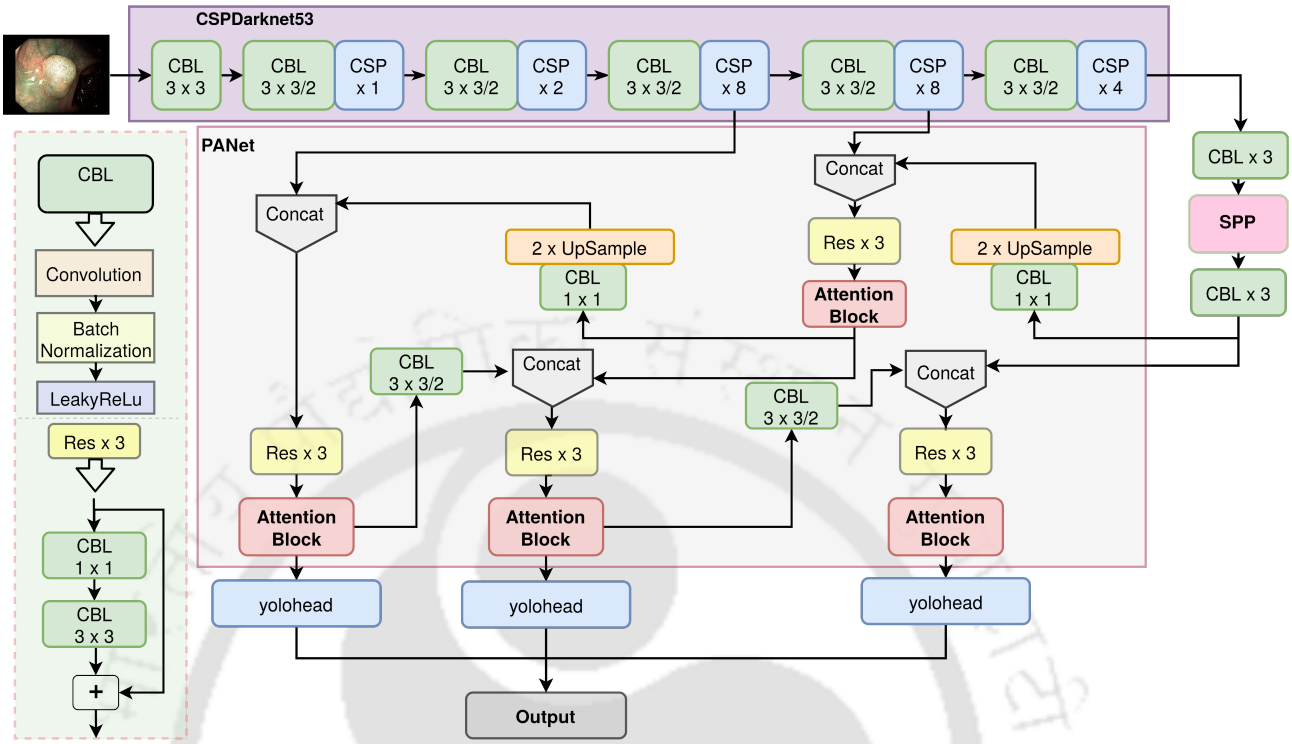


Figure 2.11: A flow of polyp classification framework using our proposed detection algorithm.

speed. It integrates all the efficient approaches which are employed in different domains. The current study discusses the ability of such models in polyp detection from colonoscopy images. Recent works on the YOLO frameworks for polyp detection show outstanding robustness and efficiency. A real-time polyp detection system by scaling the YOLOv4 algorithm is proposed by Pacal et al. [182]. They introduced the Mish activation function, DIoU loss function, and transformer blocks to the YOLOv4 architecture. They employed post-processing methods to increase the speed and accuracy of the model. However, to make the polyp detection system more generalized, Pacal et al. [183] tweaked the YOLOv3 and YOLOv4 networks by integrating Cross Stage Partial Network (CSPNet) and adopted advanced data augmentation techniques and transfer learning on big polyp datasets. The performance of polyp detection was further improved by using negative polyp samples during the training of the models. Their study also shows the effect of activation functions and loss functions of their proposed YOLO frameworks on polyp detection performances.

However, a lack of annotated data may overfit the YOLOv4 model and make it less efficient in polyp detection. Therefore, some changes corresponding to the polyp characteristics of endoscopic video frames are made in the existing model for better performance. The colonoscopy frames may

have occlusion, clutter, blur, specular noise, etc., which can degrade the detection performances. Even though HD cameras are used to acquire colonoscopy images nowadays, a small percentage of the images still suffer from such limitations. Also, the bounding box (BBoxes) used to localize the target objects may fit the arbitrary contour of objects. Therefore, various methods are generally adopted to highlight the actual target object neglecting the background. The attention mechanism is among the solutions to these problems by enabling the network to focus more on the target object. Attention mechanisms are coupled in deep detection models to learn key features of the object. It mimics the property of the human visual system. Recently, the attention mechanism has shown promising performances in various computer vision applications [184–187]. Therefore, the attention module is embedded into the backbone of CSP Darknet to focus more on the ROI of feature maps. This module would enable extraction of the polyp regions' important features, ignoring the non-polyp regions of colonoscopy frames. Our method proposes two attention modules, namely, the channel and spatial attention modules, which are incorporated into the YOLOv4. YOLOv4 extracts feature maps to three different branches to obtain three feature grid maps with various scales for detecting objects of different sizes. The three YOLO heads are then trying to localize the objects with the BBoxes. Our proposed attention modules are integrated into the feature maps before the three YOLO heads can detect and localize polyps.

2.3.3.2 Channel attention block

The channel attention block is proposed to integrate the interaction among the inter-channel feature maps. It is employed to enhance the vital information of a feature map of an object.

Let the input feature map be represented as $M \in R^{H\times W\times C}$, where H, W, and C represent the height, width, and depth of a feature map, respectively. As shown in Figure 2.12, a global average pooling operation is employed across all the depth maps to extract contextual information, embedded in the channel descriptor given as $I_c \in R^{1\times 1\times C}$, and the c-th element of I_c is given by:

$$i_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} m_c(i, j) \tag{2.14}$$

$I_c = [i_1, i_2, ..., i_c]$ and $M = [m_1, m_2, ..., m_c]$. Again, to further explore the inter-channel nonlinear relationship among the channel maps, we employ a 2-layers CNN followed by a sigmoid activation function. In order to reduce some parameters overhead, W_1 is used as the dimensionality reduction layer with a reduction factor of 16 [186]. Similarly, W_2 is used to increase the dimensionality again.

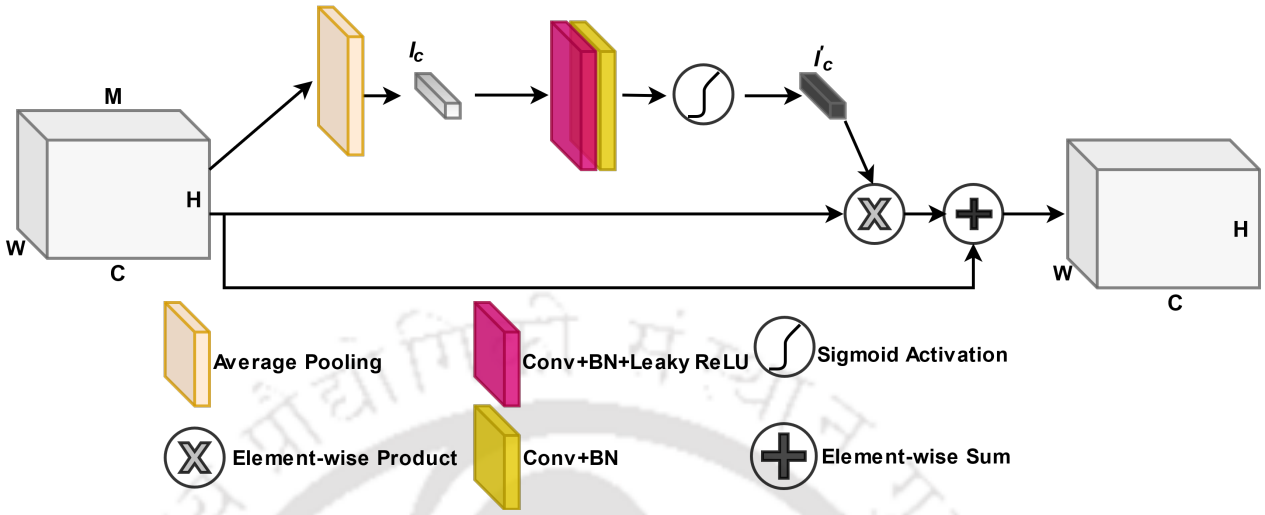


Figure 2.12: Channel attention block; Conv: Convolution, BN: batch normalization.

This process is given as:

$$I_c^{'} = \sigma(W_2\phi(W_1 I_c)) \tag{2.15}$$

where, $W_1 \in \frac{C}{16} \times C$ and $W_2 \in C \times \frac{C}{16}$.

Finally, an element-wise summation operation is adopted between the input feature map and the generated channel attention map through residual connection to mitigate the incurred information loss. The final feature map is given as: $I_c^{'} \times M + M$. The channel attention module is illustrated in Figure 2.12.

2.3.3.3 Spatial attention

The channel attention focuses on what is important in a given image. On the contrary, the spatial attention block learns where to concentrate for polyp detection. Therefore, the spatial attention module extracts the complementary features compared to the channel attention. The spatial attention mechanism focuses on the local regions of a feature map. Thus, this module is employed to preserve the local polyp ROI information in the feature maps. We initially used average-pooling and max-pooling operations along the channel axis and concatenated them to build an efficient feature descriptor to compute spatial attention. To construct a spatial attention map, we apply a convolution layer to the concatenated feature descriptor. Figure 2.13 depicts a spatial attention module. As shown in Figure 2.13, a 7 × 7 convolutional layer is introduced to aggregate the interspatial interaction of maps to

produce a one-dimensional spatial descriptor.

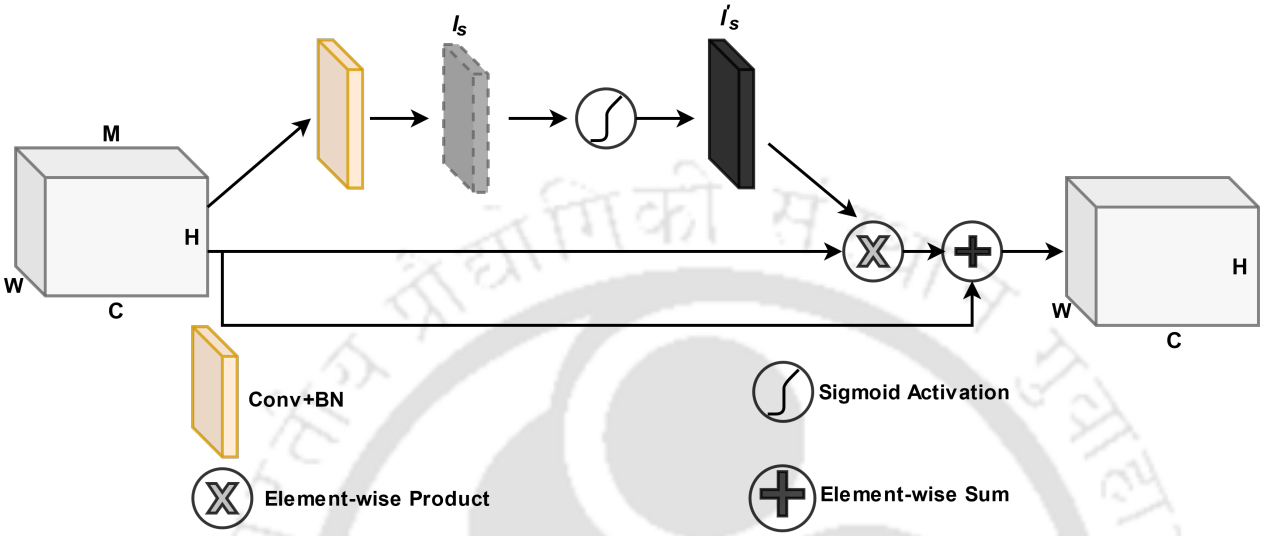


Figure 2.13: Spatial attention block.

Let the input feature map be represented as: $M \in R^{H \times W \times C}$.

Then, the generated feature descriptor I_s is represented as:

$$I_s = conv^{7\times7}(M) \tag{2.16}$$

where, $I_s \in R^{H \times W \times 1}$ and $conv^{7\times7}(\cdot)$ denotes a 7×7 convolutional layer. The sigmoid function then activates this feature map to highlight the important regions. Subsequently, it is multiplied and summed up with the input feature maps to produce the final feature map, which is given as:

$I'_s \times M + M$, where $I'_s = \sigma(I_s)$.

The detailed experimental setup, training, and performances are discussed in the experimental results section.

2.3.4 Results and discussion

2.3.4.1 Evaluation metrices

In this work, some of the extensively recommended standard metrics are used to evaluate detection and localization performances. [71, 188].

- $IoU(A, B) = \frac{A \cap B}{A \cup B}$, measures the overlap between two bounding boxes A and B as the ratio between the overlapped area.

- AP: Average precision was computed as an average APs for IoU from 0.25 to 0.75 with a step-size of 0.05.
- FPS=$\frac{\#frames}{sec}$.

True positive (TP), False positive (FP), False Negative (FN), and True Negative (TN) are essential to finding other metrics, helpful to evaluate performances. Considering the problem of polyp detection, these values can be considered as follows. TP indicates that the predicted bounding box falls on the ground truth of the polyp. FP suggests that the predicted bounding box falls outside the ground truth of the polyp. FN shows no predicted bounding box but a polyp of the frame. Finally, TN indicates that no polyps are detected in images without polyps. Precision ($Pre.$) indicates how many of the values estimated as positive are actually positive. Recall ($Rec.$) indicates how many true positives are estimated from all positives.

$$\text{Precision} = \frac{TP}{TP+FP},\ \text{Recall} = \frac{TP}{TP+FN},\ F1 = 2 * \frac{Pre. \times Rec.}{Pre. + Rec.},\ F2 = 5 * \frac{Pre. \times Rec.}{4 * Pre. + Rec.} \quad (2.17)$$

2.3.4.2 Experimental setup and configuration

Two databases are used in our experiment for this study. The details of the databases are given in Table 2.7. The Kvasir-SEG dataset [157] contains 1000 polyp images and their corresponding ground truth from the Kvasir Dataset v2. The images and their corresponding masks (The bounding box (coordinate points) for the corresponding images) are available. According to the usability of the data statement, it is mentioned that the data can be suitable for general segmentation and bounding box detection research. We used 800 images for training and the remaining 200 images for the validation in a five-fold cross-validation manner. Similarly, the SUN Colonoscopy Video Database [158] is developed for the evaluation of automated colorectal polyp detection systems. The images are labeled by expert endoscopists. This dataset contains 49,136 polyp frames from 100 patients with fully annotated polyps with bounding boxes and 109,554 frames without polyps. The groundtruth information for polyp frames are available in the form of bounding box coordinates. Therefore, in order to make non-overlapping sets for training and test sets, we used first 80 patient's samples for training and the rest images from 20 patients were kept for testing. The total number of images available for training and testing was 40,707 and 8429, respectively. Thereby we could make the training and test ratio approximately 80:20. The sizes of the images were made 416 × 416. The model was tested

in Google Colab (cloud GPU) with Nvidia Tesla T4 @585 MHz. The hyperparameters set for the YOLOv4+Attention model are as follows: Learning rate: $1e^{-3}$, batch size: 64, anchors: 3, and threshold: 0.25.

2.3.4.3 Polyp detection performance

Table 2.8 shows the polyp detection and localization performances by different baseline models on the Kvsir-SEG dataset. It can be observed that our method achieves an average precision (AP) of 0.8971, which is the best among all. The APs achieved at multiple IoU thresholds i.e $\mathbf{AP_{25}}$, $\mathbf{AP_{50}}$, and $\mathbf{AP_{75}}$ are 0.9485, 0.9279, and 0.7849, respectively. The IoU measures the precision at which the bounding box localizes the target object. Our results clearly show that our method is better at localizing polyp ROIs than the state-of-the-art deep architectures often used in medical image analysis. A few available methods have used the Kvasir-SEG dataset to evaluate the polyp detection algorithms. The performance of our proposed method is compared with the current state-of-the-art methods and is presented in Table 2.9.

Table 2.8: Detection performace of baseline models on the Kvsir-SEG dataset.

Method	Backbone	AP	IoU	$\mathbf{AP_{25}}$	$\mathbf{AP_{50}}$	$\mathbf{AP_{75}}$	FPS
EfficientDet-D0 [189]	EfficientNet-b0	0.4756	0.4322	0.6846	0.5047	0.2280	35.00
Faster R-CNN [190]	ResNet50	0.7866	0.5621	0.8947	0.8418	0.5660	8.00
RetinaNet [191]	ResNet50	0.8697	0.7313	0.9395	0.9095	0.6967	16.20
RetinaNet [191]	ResNet101	0.8745	0.7579	0.9483	0.9095	0.7132	16.80
YOLOv3+spp [180]	Darknet53	0.8105	0.8258	0.8856	0.8532	0.7586	45.01
YOLOv4 [181]	Darknet53, CSP	0.8513	0.8025	0.9348	0.9128	0.7757	66.67
YOLOv4+Attention	Darknet53, CSP, Attention	**0.8971**	**0.8325**	**0.9485**	**0.9279**	**0.7849**	50

Table 2.9: Comparision of performances with the state-of-the-art methods on the Kvsir-SEG dataset.

Method	Pre.	Rec.	F1	AP	IoU	$\mathbf{AP_{25}}$	$\mathbf{AP_{50}}$	$\mathbf{AP_{75}}$	FPS
Attentive YOLOv5 [192]	0.915	0.899	0.907	-	-		-	-	35.71
ColonSegNet [71]	-	-	-	0.8000	0.8100	0.9000	0.8166	0.6706	180
Proposed	0.9324	0.8457	0.8869	**0.8971**	**0.8325**	**0.9485**	**0.9279**	**0.7849**	50

Figure 2.14 shows qualitative results of some samples from the Kvsir-SEG dataset for polyp detection. Results from the recent state-of-the-art method, YOLOv4, and our proposed method are shown in Figure 2.14. From the figures, it can be observed that both YOLOv4 and our proposed method

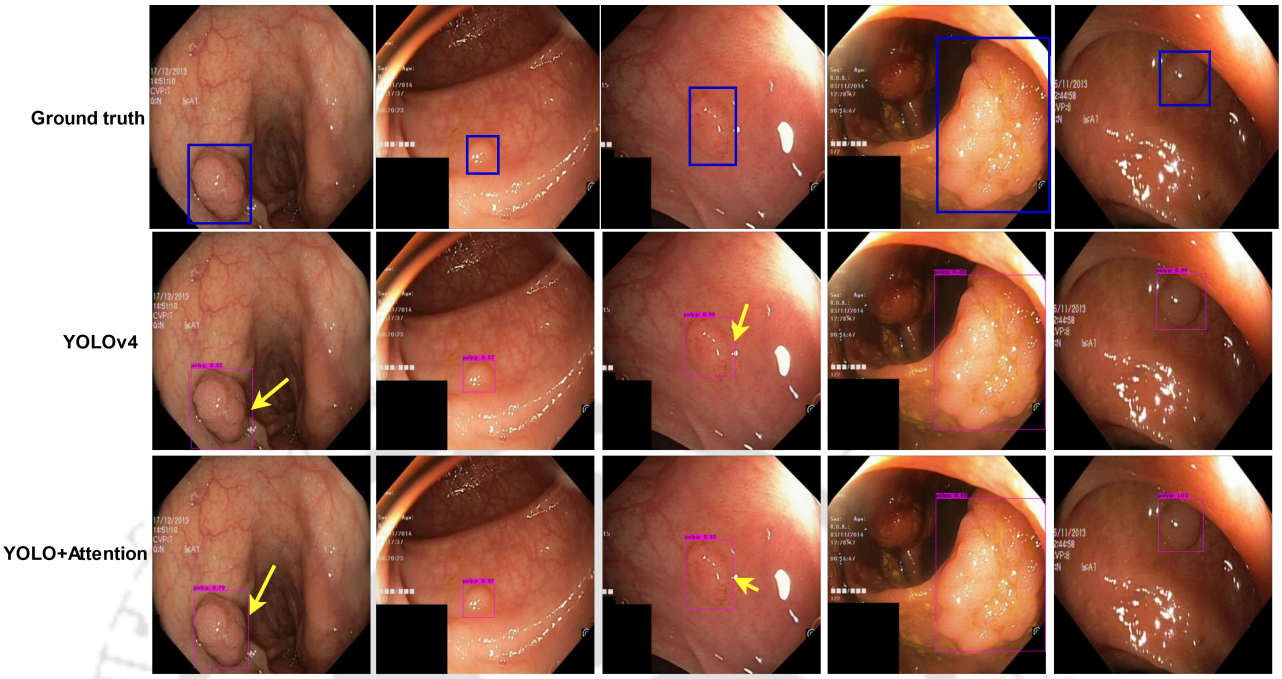


Figure 2.14: Detection and localization results on test dataset: Kvsir-SEG.

can detect and localize polyps with high confidence. Some of the bounding boxes are annotated with the yellow arrows to show that our proposed method is better in localizing the polyps. In YOLOv4, most polyps are localized with wider bounding boxes than the proposed Attention YOLOv4 model. This is also validated by the quantitative results, where the average IoU for the Attention YOLOv4 is better than YOLOv4, as shown in Table 2.8 and Table 2.9. Samples with the blue bounding boxes are ground truths and are available with the dataset.

Table 2.10: Detection performace of our proposed method compared to the state-of-the-art methods on the SUN Colonoscopy Video Database.

Method	**Pre.**	**Rec.**	**F1**	**F2**	**AP**	**IoU**	**AP_{25}**	**AP_{50}**	**AP_{75}**	**FPS**
YOLOv4-CSP	0.9387	0.8325	0.8824	0.8517	0.8597	0.8052	0.9762	0.9621	0.6408	66.67
YOLOv4 +negative samples [183]	0.9379	0.8500	0.8918	-	-	-	-	0.9845	-	-
Proposed	0.9425	0.8274	0.8812	0.8481	0.9172	0.8179	0.9868	0.9721	0.7328	50

The performances on the SUN database with YOLOv4 and the proposed method are also shown in Table 2.10. The qualitative localization performances on some of the samples of the SUN database are shown in Figure 2.15. It is observed that similar performances are also achieved on this dataset. YOLOv4 model did not detect the second image of the second-row polyp, but our proposed model

Table 2.11: Cross dataset detection and localization performace: Trained on SUN Colonoscopy database and tested on Kvsir-SEG dataset.

Method	Backbone	AP	IoU	AP_{25}	AP_{50}	AP_{75}	FPS
YOLOv4	Darknet53, CSP	0.8597	0.7240	0.8789	0.7945	0.5287	66.67
YOLOv4+Attention	Darknet53, CSP, Attention	**0.9172**	**0.7667**	**0.9231**	**0.8600**	**0.6144**	**50**

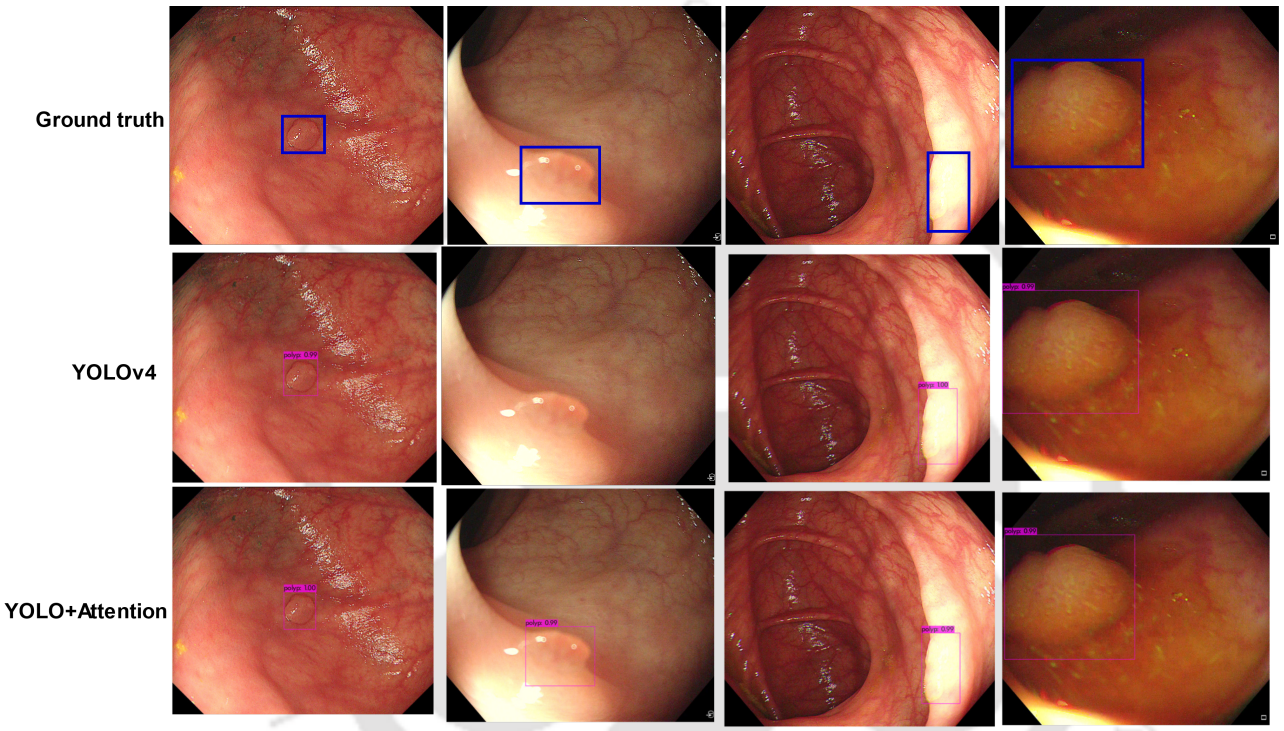


Figure 2.15: Detection and localization results on test dataset: SUN Colonoscopy database.

could detect and localize it. Further to validate the robustness of our model, we also cross-validated the performances. We evaluated the performance using the test data from the Kvasir-SEG dataset while the model was trained with the SUN database. The performance on the cross dataset is shown in Table 2.11.

The efficiency of the algorithm in detecting polyps is also quantified by the IoU scores, which represent how compactly the bounding boxes can localize the polyps. The state-of-the-art method YOLOv4 has an IoU of 80.25%, whereas our method gives an IoU score of 83.25%. This validates the fact that our proposed method could localize the polyps perfectly. Furthermore, during the analysis of SUN colonoscopy videos, our algorithm detected not only all the polyps that were found by endoscopists but also additional polyps that were not detected by the endoscopists. These findings

suggest that our automatic polyp detection algorithm is practical and accurate. The algorithm's feasibility in real-world clinical practice is also supported by its short processing time. Because our algorithm can process images at a speed of 50 fps for polyp detection, it can be employed for real-world colonoscopy as colonoscopy video encodings usually have standardized rates of approximately 30 fps.

The two databases used for validating our proposed method for polyp detection suggest that the algorithm could detect various polyp structures quite accurately. It is crucial in clinical practice as early detection helps better prognosis and clinical management, leading to a higher survival rate. This study has several limitations. First, there could be selection bias in the training and test dataset. However, we believe that splitting the database was done such that the training and test datasets contain non-overlapping images. It did not deviate from the quality of usual real-time colonoscopy videos in daily practices. Second, the state-of-the-art YOLOv4 method showed an AP and IoU of 85.13% and 80.25% on the Kvsir-SEG database. Similarly, on the SUN Colonoascopy Video Dataset, it shows the AP and IoU of 85.97% and 80.52%. We could increase polyp detection performances in both databases by introducing attention blocks in the YOLOv4 framework. Though the attention mechanism helps improve the detection rate, other attention could be coupled for further improvement. We suggest that additional training with a more considerable amount of training data may further enhance the performances. The degradation of performances by our algorithm can be attributed to colonoscopy features such as bubbles, specularity, normal mucosa , blood traces, and polypectomy site, which visually look similar to the polyps. Third, all the images were obtained using the HD Olympus endoscope system. Thus, the performance of our algorithm to other imaging modalities is still to be validated. However, we believe the algorithm may function with other endoscope systems after fine-tuning and domain adaptation. Finally, we analyzed recorded videos rather than real-time colonoscopies, limiting the applicability of our algorithm in daily clinical practice. Nonetheless, our algorithm can be applied to real-world colonoscopy procedures because of the short processing time and high performance for unaltered videos, which theoretically represent real-time colonoscopy. Furthermore, we are confident that the applicability of our algorithm to real-time colonoscopy is supported by our meticulous validation, which involved different datasets. Also, the cross-validation performance of our algorithm provides the relevance of our algorithm in real-time analysis.

2.3.5 Conclusion

This work presents a deep attention-based YOLOv4 framework to detect polyps in colonoscopy images. The attention module in the YOLOv4 encapsulates spatial and contextual information of the polyp ROIs effectively. The attention module selectively accentuates the polyp ROIs by extracting local and global information. The performance of the suggested algorithm outperforms state-of-the-art approaches by a significant margin. The consistency of results across datasets also demonstrates the generalizability and robustness of our method. We hope to improve polyp detection in the future by training the network with features that best characterize the clinical manifestations exhibited by the polyps. Adaptive attention and domain adaptation could be coupled in the frameworks for better generalizability and efficiency for polyp detection. We would like to validate our model exclusively on the sessile and diminutive polyp datasets as they are often missed. This work presents the framework for detecting polyps using bounding box localization. However, sometimes perfect delineation of poly ROIs is essential and can have considerable clinical significance. Therefore, in the next chapter, we will discuss the importance of polyp segmentation and will try to develop methods for polyp segmentation frameworks. This would be another way of looking at polyp detection from the frames by perfectly segmenting the polyp.

2.4 Summary

The two methods discussed in Chapter 1 propose polyp detection frameworks. The first approach uses color, texture, and shape information of the polyps to localize them in the frames. This approach is more suitable for the offline processing of the colonoscopy videos. Also, the proposed method sometimes misses polyps that are textureless and do not show any significant clinical manifestations to the naked eyes. Therefore, the objective of our subsequent work was to present a strategy that focuses more on the polyp ROIs and learn the hidden discriminating features from the non-polyp regions. Thus, our second approach proposes using spatial and channel attention in the YOLOv4 framework to extract spatial and contextual information of polyps. This method could able to detect and localize the small and serrated polyps. The generalization and robustness of the method can be validated from the extensive qualitative and quantitative results on datasets having a variety of polyp structures. This approach can also do real-time polyp detection. The localized polyps are used to extract features to classify them into different grades of carcinoma. However, the bounding box

around the polyp ROIs may introduce significant non-polyp tissues, which may harm the performance of a classifier. The non-polyp tissues introduce unnecessary features in the feature extraction stage. Also, the processing of extra pixels may affect the computational complexity of the classifier.

3

Polyp Segmentation

Contents

Objective

This chapter proposes techniques to delineate polyps in colonoscopy videos. Polyp segmentation in colonoscopy video frames is an essential step for effective diagnosis. It is a way of detecting polyps perfectly by segmenting exact polyp boundaries. In this view, our first work is attempted to segment polyps from the polyp frames needing immediate attention. Recognizing such polyp frames is achieved by utilizing the depth information of the polyps in this work. The depth information is quite important for polyp analysis. Therefore, the first stage of this method extracts the clinically significant frames (key-frames) from a video by extracting depth and other information. Further, the prominent and elevated polyps which are at higher risk of being cancerous are segmented from the selected key-frames. However, small and serrated polyps that may develop malignancy over time must be detected early. Therefore, our second approach tries to segment any polyp structures using color and texture features of the polyps in an adaptive Markov random fields (MRF) framework that effectively preserves contextual and spatial information. However, polyps may not show any color and texture characteristics in their nascent stage. To effectively recognize such polyps, our third approach uses the polyp shape information on a saliency map-based approach to segment polyps effectively. The deep-learning-based U-Net architecture is used to generate the saliency map. The segmented polyps can further be utilized to extract useful features for their classification.

3.1 Introduction

During the colonoscopy, doctors comprehensively analyze the detected polyp regions to find dysplasia in them. Depending on the nature of the polyps, they may opt for laparoscopic surgery. However, the number of frames captured during the entire colonoscopy process is so humongous that it challenges the surgeon to infer useful clinical information. Therefore, video summarization techniques are adopted that only retain the clinically informative frames. During WCE, the capsule moves under the peristalsis movement, and it is challenging to control the motion and orientation of the camera. Thus, redundant and clinically non-significant frames are generally obtained in a video sequence. WCE takes nearly 8 hours, capturing close to 50000 frames. A large part of the data is clinically not significant and needs to be removed [10]. Selecting the frames having sufficient clinical information reduces the burden on the clinicians from reviewing uninformative frames. Also, the key-frames can provide better features, thus helping in effective diagnosis. The unnecessary polyp features extracted from the bad frames

may degrade the performance of an automated diagnosis system. Therefore, this step is preferred and adopted before segmenting the polyps. Preliminary work on key-frame selection and subsequent polyp segmentation is proposed as the first approach in this chapter. The other two techniques presented in this chapter are aimed at effective polyp segmentation from the already acquired significant polyp frames. The rest of the chapter is organized as follows. The first approach based on polyp depth information is presented in section 3.2. The second approach is based on the spatial and contextual information of the polyp structures and is described in section 3.3. Section 3.4 presents the third approach which is based on the saliency map and shape information of the polyps. The summary of the chapter is provided in section 3.5

3.2 Key-Frames and Segmentation Using Depth Information

A recent work focusing on video summarization instead of anomalies detection like bleeding or ulceration is proposed by Li et al. [149]. Iakovidis et al. [150] used clustering-based methods for video summarization. Similar work based on the clustering technique was proposed by Avila et al. [151]. However, clustering-based methods are not suitable in noise environments. Endoscopy frames are generally susceptible to noise. Also, redundant frames are captured during the endoscopy, which makes clustering methods perform poorly. Researchers are working on visual attention models, like saliency maps, for finding key-frames of videos [152]. Another visual saliency-based attention model was proposed by Ezaj et al. [153]. They used motion, color, and texture features for hysteroscopy video summarization. A color histogram comparison-based method was adopted by Mendi et al. [154]. They compared the color histogram of successive frames in a video sequence, and key-frames were selected using k-means and PCA whenever a significant change in content was observed. However, this model does not fit into endoscopic videos as most of the frames have similar color information. Recently, dictionary learning-based approaches have been proposed for video summarization [155]. In [156], a gastroscopic video summarization technique based on a dictionary learning approach is proposed.

Key-frames are very important and help in better prognosis and clinical management of the disease. Therefore, colonoscopy frames that need immediate medical attention are considered for this study. Malignant polyps usually have a convex shape and are more textured compared to benign polyps. Seitz et al. [193] proposed that polyp size is correlated to the degree of dysplasia. A large and convex type polyp is associated with more severity of dysplasia. Getting a 3D view of the polyp surface can

significantly help in resection [194]. A good 3D reconstruction of an object in an image entails dense depth estimation. The 3D view gives shape and size information of a polyp. Depth estimation of endoscopic images is challenging as the endoscopic images are monocular.

Attempts have been made to solve it as a per-pixel regression problem, however, supervised learning methods require a lot of training data. It isn't easy to acquire depth data without using stereo cameras or expensive depth sensors, as with endoscopy videos. Thus unsupervised methods are being given more importance. Depth estimation in endoscopic video frames imparts clinical relevance to a physician. 3D reconstruction of the monocular images helps in diagnosis and surgical planning. Recently, depth estimation, especially monocular depth estimation (MDE) has gained high research interest. This is due to its application in scene understanding, robotics, autonomous driving, and Augmented Reality (AR). Finding depth from a single image is an unconstrained problem since many real-world scenes can give the same 2D image, resulting in the same depth maps. Humans perceive depth from cues such as perspective, prior knowledge of sizes of objects, or occlusion. Both supervised, and unsupervised-based methods have been employed in the literature for estimating depth.

Eigen et al. [195] introduced a multiscale information approach that takes care of both global scene structure and local neighboring pixel information. A scale-invariant loss is used for MDE. Similarly, Xu et al. [196] formulated MDE as a continuous random field problem (CRF). They fused the multi-scale estimation computed from the inner semantic layers of a CNN with a CRF framework. Instead of finding continuous depth maps, Fu et al. [197] estimated depth using an ordinal regression approach. A space-increasing discretization method is introduced by allowing objects at larger depths to have a lesser influence on the depth maps than the objects nearer to the observer.

Depth is generally obtained using sensors like LIDAR, Kinect, or by using stereo cameras. Sensors are expensive, and stereo cameras are not generally used in endoscopy due to several restrictions. Obtaining ground-truth training data for depth estimation is very difficult in endoscopic imaging, so supervised methods are not feasible for endoscopic image classification. Finding correspondence between two images for 3D reconstruction is also difficult in endoscopy videos. It isn't easy to find corresponding features across the frames.

Hence, unsupervised and semi-supervised methods are employed for MDE. Garg et al. [198] used binocular stereo image pairs for the training of CNNs and then minimized a loss function formed by the wrapping of the left view image into its right of the stereo pair. Godard et al. [199] improved

this method by using the left-right consistency criterion. They trained CNNs on stereo images but used a single image for inference. They introduced a new CNN architecture that computes end-to-end MDE. The network was trained with an efficient reconstruction loss function. The state-of-the-art unsupervised MDE method, i.e., Monodepth [199] model has limited application in in-vivo images like endoscopic images. This is because most models leverage outdoor scenes [200] and a few indoor scenes [201] for training, and they use high-end sensors or stereo cameras, while the WCE method only captures monocular images. Hence, it is important to devise a strategy to perform MDE in medical imaging datasets that generally do not have ground truth depth information. That is why a transfer learning approach is adopted in our method for estimating depth. Transfer learning refers to a learning method where what has been learned in one setting is exploited to improve generalization in another setting [202]. Zero-shot learning is the extreme case of transfer learning where no labeled examples are present. In our method, a zero-shot learning approach for MDE [203] is employed.

The proposed method[1] consists of two main steps. The first step focuses on depth estimation, and the second step extracts key-frames. As mentioned above, a zero-shot learning approach is adopted for depth estimation in endoscopic videos. We propose a framework to select the most informative frames of an endoscopic video sequence. Our method employs a three-criteria approach to identify the key-frames. Subsequently, the key-frames can be used for 3D reconstruction. Our method is unique in the sense that it considers depth information to find key-frames. Experimental results clearly demonstrate the effectiveness of our method in choosing the key-frames and subsequent polyp visualization and segmentation. The proposed methodology is elucidated in section 3.2.1. Experimental results and conclusions are discussed in section 3.2.2 and section 3.2.3, respectively.

3.2.1 Proposed method

The polyps' depth is extracted using a transfer learning-based approach. The key-frames are selected using the extracted depth information and other image properties of the colonoscopy frames. The steps of the proposed approach are described below.

3.2.1.1 Depth estimation

Due to the unavailability of ground truth depth data in endoscopy video datasets, a transfer learning approach is adopted for MDE in our proposed method. Lasinger et al. [203] proposed a zero-shot

[1]This work has been published in IEEE Access, 2021 (Refer *List of publications* page for details).

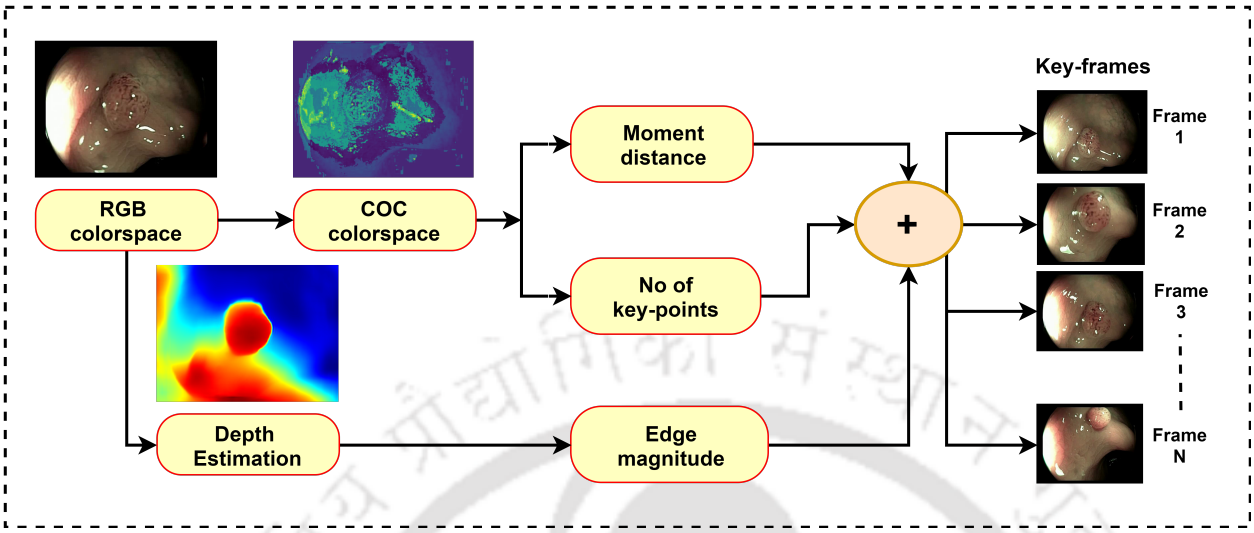


Figure 3.1: Proposed method of finding key-frames.

learning for depth estimation. The work of Lasinger et al. inspires our proposed work for depth estimation as a zero-shot approach. The flow diagram showing the proposed method of finding key-frames using depth information is shown in Figure 3.1.

This section explains how we use monocular images to learn relative depth. As demonstrated in Figure 3.2, we model monocular relative depth perception as a regression problem. In an end-to-end method to regress pixel-wise relative depth given a batch of input images I, we create a non-linear function $y = f(I, \delta)$ parameterized by δ. The network is built on a feedforward ResNet architecture that generates multi-scale feature mappings [142]. To improve predictions, a progressive refinement technique is used to combine multi-scale variables.

The model was trained for depth maps obtained in three different ways. First, the dataset contains depth maps obtained using LIDAR sensors. This method gives depth maps of high quality. Second, the Structure from Motion (SfM) approach is employed to estimate the depth. The third method of getting depth information from stereo images of the 3D movies dataset. It uses optical flow to find motion vectors from each of the stereo images. Then, the left-right image disparity is used to find a depth map. The dataset contains images that have varying aspect ratios. Sometimes, black bars on frame borders appear in estimated depth maps. So, all the images are cropped to extract only the center portion of the frame. This ensures the framework can handle images of varying aspect ratios. Moreover, the method focuses more on the central part of the image frame. Using the distance of an object from the camera to predict depth leads to sparse 3D reconstructions. This is because depth is

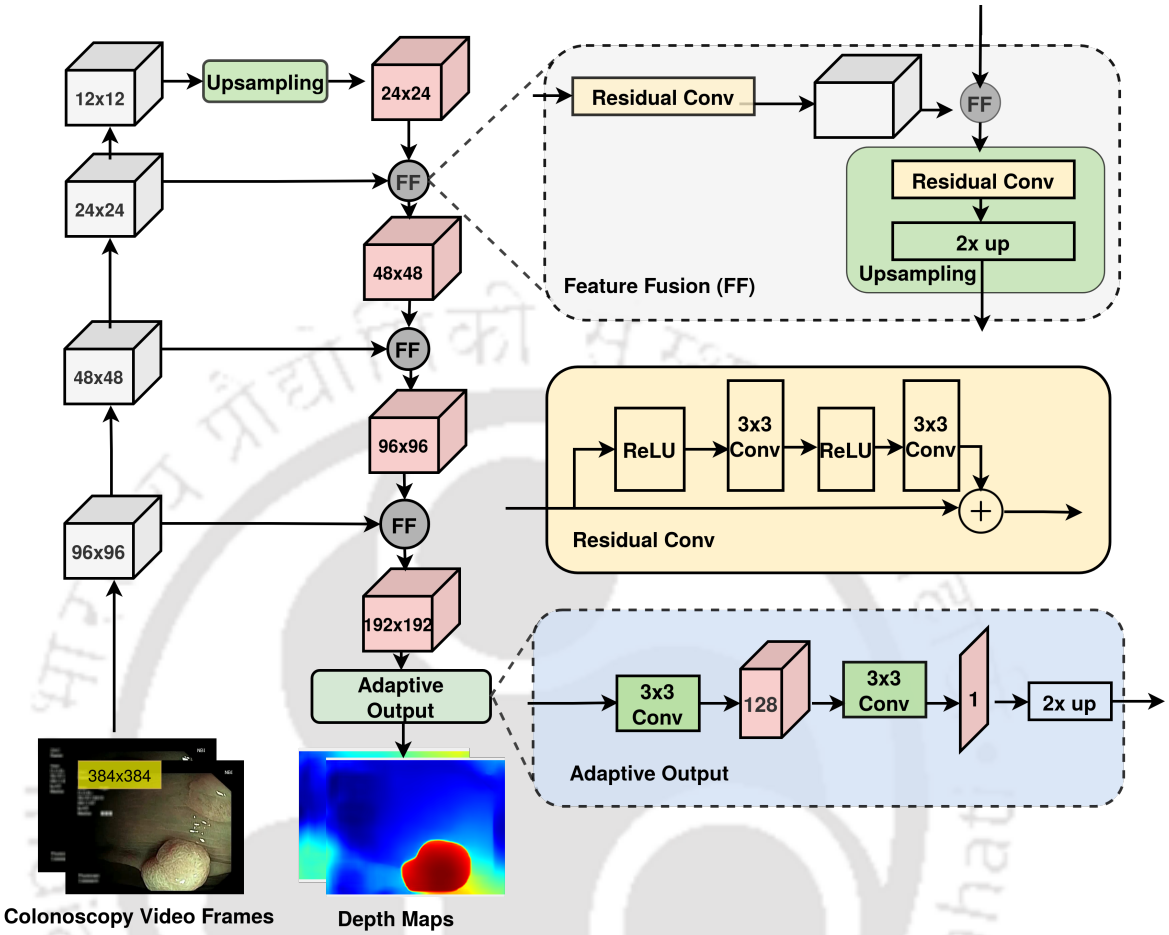


Figure 3.2: Network architecture for Depth Estimation from colonoscopy video frames; The model is based on a feedforward ResNet architecture.

estimated by tracking the corresponding features over a series of frames. Then, the induced parallax is used for triangulation and depth estimation. However, the resultant parallax will be small for distant features (like the sky) and won't allow proper reconstruction. Thus, distant objects like the sky are not considered while estimating depth. This addresses the issue of finding correspondences for distant objects.

The disparity map is found by using stereo matching using optical flow. Optical flow successfully handles moderate displacements. The horizontal component of the flow vectors is used as a reference for finding a disparity map. Optical flow is estimated by taking either the left or right image as a reference and finding flow from the other. Next, the consistency between both left and right is calculated to discard the pixels with more than one-pixel disparity.

The datasets on which the model is trained are unique because they contain both positive and neg-

ative disparities. However, training on ground truth data from different sources has some constraints: 1) The dataset contains images that have only depth (from LIDAR sensors) or disparity images; 2) Data obtained from the SfM technique gives depth images for which scale is not known; 3) The 3D movies dataset gives a ground truth depth which has an unknown shift.

Loss function. A shift and scale-invariant loss function is chosen to address the problems pertaining to training on three different datasets. Let $\mathbf{d} \in \mathbb{R}^N$ be the computed inverse depth and $\mathbf{d}' \in \mathbb{R}^N$ be the ground truth inverse depth, where N is the number of pixels in a frame. Here s and t represent scale and shift, respectively and, they are positive real numbers. This can be represented in a vector form by taking $\vec{\mathbf{d}}_i=(\mathbf{d}_i, 1)^\intercal$ and $\mathbf{p}=(s,t)^\intercal$, and thus, the loss function becomes:

$$\mathcal{L} = \arg\min_{s,t} \quad \frac{1}{2N}\sum_{i=1}^{N}(s\mathbf{d}_i + t - \mathbf{d'}_i)^2 \tag{3.1}$$

$$\mathcal{L}(\mathbf{d}_i, \mathbf{d'}_i) = \arg\min_{\mathbf{p}} \quad \frac{1}{2N}\sum_{i=1}^{N}(\vec{\mathbf{d}}_i^\intercal\mathbf{p} - \mathbf{d'}_i)^2 \tag{3.2}$$

The closed-form solution is given as:

$$\mathbf{p}^{opt} = (\sum_{i=1}^{N}\vec{\mathbf{d}}_i\vec{\mathbf{d}}_i^\intercal)^{-1}(\sum_{i=1}^{N}\vec{\mathbf{d}}_i\mathbf{d'}_i) \tag{3.3}$$

Substituting $\mathbf{p}^{opt}$ into (3.2) we get:

$$\mathcal{L}(\mathbf{d}_i, \mathbf{d'}_i) = \arg\min_{\mathbf{p}} \quad \frac{1}{2N}\sum_{i=1}^{N}(\vec{\mathbf{d}}_i^\intercal\mathbf{p}^{opt} - \mathbf{d'}_i)^2 \tag{3.4}$$

Regularization term. A multi-scale scale-invariant regularization term is used, which does gradient matching to the depth inverse space. This biases discontinuities to be sharp and coincide with ground truth discontinuities. The regularization term can be defined as,

$$\mathcal{L}_r(\mathbf{d}_i, \mathbf{d'}_i) = \frac{1}{N}\sum_{j=1}^{k}\sum_{i=1}^{N}(|\Delta_x {Q_i}^k| + |\Delta_y {Q_i}^k|) \tag{3.5}$$

where,

$$Q_i = \vec{\mathbf{d}}_i^\intercal\mathbf{p}^{opt} - \mathbf{d'}_i \tag{3.6}$$

Here Q^k gives the difference of inverse depth maps at a scale k. We use $k = 4$ scale levels, halving the image resolution at each level. Also, the scale is applied before finding x and y gradients.

Modified loss function. The final loss function for a training set of size M, taking into consideration

of the regularization term, becomes:

$$\mathcal{L}_{final} = \frac{1}{M}\sum_{i=1}^{M}\mathcal{L}(\mathbf{d}^i, (\mathbf{d}')^i) + \alpha\mathcal{L}_r(\mathbf{d}^i, (\mathbf{d}')^i) \tag{3.7}$$

Here α is taken as 0.5.

3.2.1.2 Selection of key-frames

During the colonoscopy, not all the captured frames are clinically significant. Most of the frames may have redundant information, or may not be useful from a diagnostic perspective. Such frames need to be discarded and the clinically informative frames need to be retained. It is also strenuous and computationally intensive for a physician to investigate each frame of a video sequence. Thus, we propose a key-frame selection technique. Subsequently, 3D reconstruction is done to perform further analysis of the polyps. The key-frame selection method is given in Figure 3.1.

Colour space conversion. Our dataset contains images that are in RGB color space. Taking clues from the human visual system, which works on saliency, we changed the color space from RGB to COC, which gives a better perception in the medical imaging [204].

The image is subsequently used to find key-frames. A frame should satisfy three criteria before being selected as a key-frame: 1) It should be significantly different from neighboring frames. 2) The key-frame should give significant depth information of a polyp. 3) The polyp should not be occluded in the key-frame. We ensured that the above requirements were met, and they are formulated as follows:

Image moment: Image moments give the information of the shape of a region along with its boundaries and texture. Hu moments [205] are considered as they are invariant to affine transformation, and moment distances of consecutive frames are used to identify the redundant frames of a video. Subsequently, the moment difference between consecutive frames is calculated. The frames with a higher moment distance will be considered as a key frame. The moment distance d between two images is calculated as:

$$d = \sum_{i=1}^{i=7}(I_i - I_i')^2 \tag{3.8}$$

where, i represents each of a total of 7 moments.

Edge density: In our proposed method, the key-frames which have significant depth information

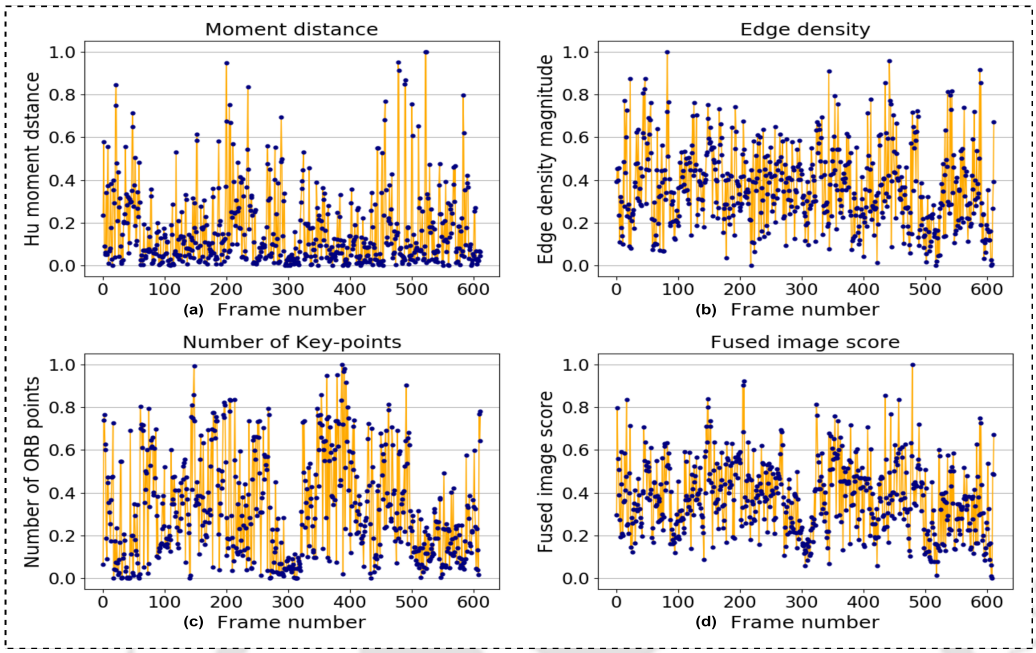


Figure 3.3: Plot of Moment distance, Edge density, Number of key-points and the total fused score vs frame number of a colonoscopy video sequence.

are only considered for the 3D reconstruction of a polyp. It is observed that the polyp images having more edges have more depth information. The edge information can be obtained with the help of the gradient magnitude of an image. Before finding the gradients, images were smoothed using a Gaussian kernel.

Horizontal and vertical gradients are obtained using Sobel operators S_x and S_y, and then the gradient magnitude ΔS is calculated as follows:

$$\Delta S = \sqrt{(S_x)^2 + (S_y)^2} \tag{3.9}$$

Key-point detection: The proposed moment-based key-frame detection method may capture some occluded frames. So, the objective is to select non-occluded key-frames from a group of key-frames that were extracted by our proposed image moment and edge density-based criteria. For this, a key-point detection-based technique is used. For key-point detection and extraction, we used ORB (Oriented FAST and Rotated BRIEF). ORB is computationally faster and robust to noises in endoscopic images. The frames containing a lesser number of ORB points correspond to occluded polyps.

Adaptive key-frame selection. After finding the moment distance (d), edge magnitude (s), and the number of ORB points (p), we normalize these scores using min-max normalization. This is

done so that each of the three scores is reduced to the range of 0 to 1 with both values inclusive. Instead of adding the three scores directly, we use dynamic weights to capture the changes in a video. The variable having more significant variance is given more weightage. Here, w_i is the weight of the normalized score. To consider intra-variable changes, we used the sum of the magnitude of difference between consecutive frame scores as a measure to find weights. We then normalized this score to be used as weights for finding a fused score. The weights are given by:

$$d_1 = \sum_{i=1}^{n} |d_i - d_i'|, s_1 = \sum_{i=1}^{n} |s_i - s_i'|, p_1 = \sum_{i=1}^{n} |p_i - p_i'| \tag{3.10}$$

$$w_1 = \frac{d_1}{d_1 + s_1 + p_1}, w_2 = \frac{s_1}{d_1 + s_1 + p_1}, w_3 = \frac{p_1}{d_1 + s_1 + p_1} \tag{3.11}$$

$$f = w_1 d_1 + w_2 s_1 + w_3 p_1 \tag{3.12}$$

here, d_1, s_1, p_1 are the sum of magnitudes of difference between consecutive frame scores and f is the fused score obtained by adaptively weighting the three frame scores. The frames with the highest fused scores are selected according to a threshold value which was set as 0.5. The variance of each criterion with frame number is shown in Figure 3.3.

3.2.2 Results and discussion

The proposed method is evaluated on the publicly available dataset. This dataset contains colonoscopic video sequences from three classes: adenoma, serrated, and hyperplasic. The adenoma class contains

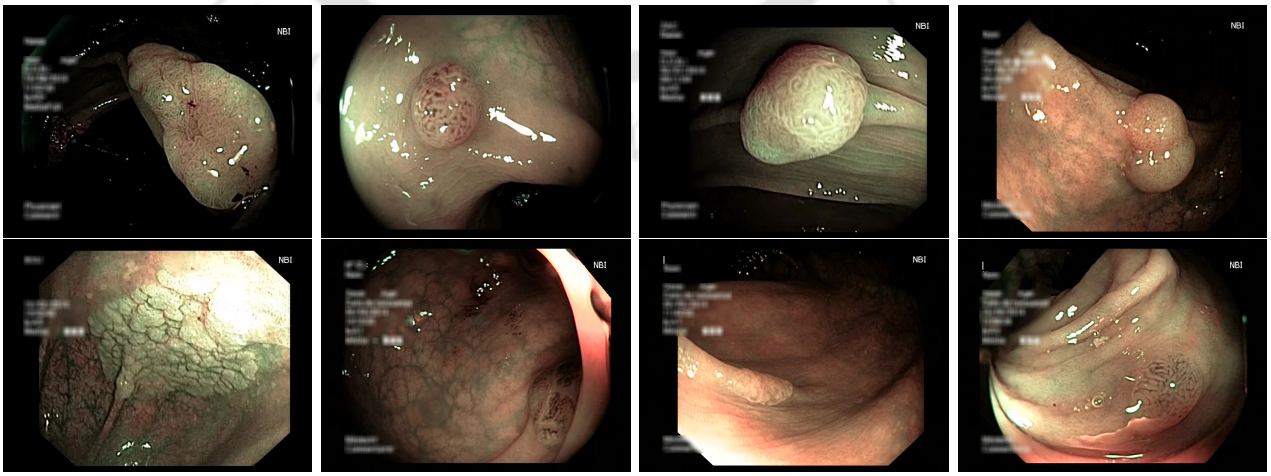


Figure 3.4: Some images of colonoscopy dataset: the first row are the examples of convex polyps and the second row are the examples of patchy polyps.

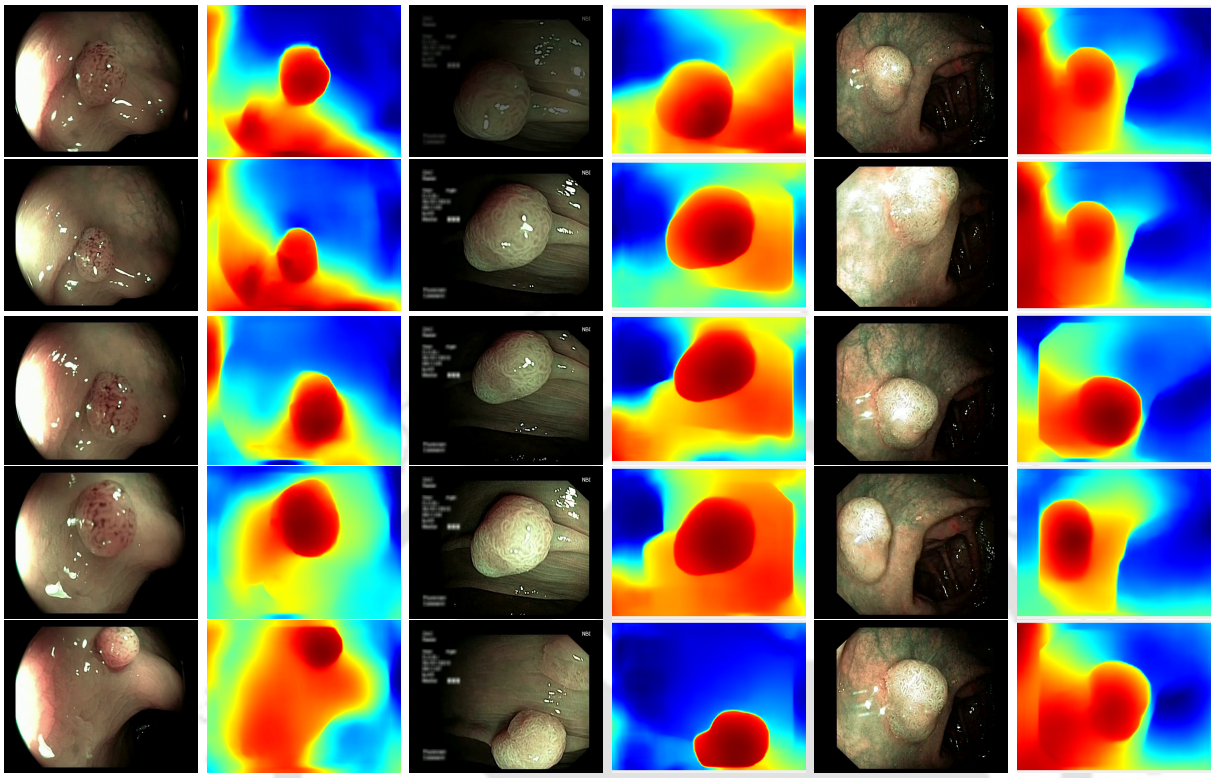


Figure 3.5: Key-frames obtained by our method and their corresponding depth maps. The polyp is visible from different viewing angles in these selected frames.

40 sequences, the serrated category contains 15, while the hyperplasic category contains 21 video sequences [128]. In this work, we consider only the frames from the adenoma (malignant) class because this class needs the maximum attention of the physician. The dataset used in this work is publicly available in the url: `http://www.depeca.uah.es/colonoscopy_dataset/`.

We considered only narrowband images (NBI) for this work as they require less preprocessing and are generally used for polyp classification. The adenoma class contains 40 video sequences of different patients. In this work, the frames with convex polyps are taken to estimate the depth. A few convex and patchy polyp images of the dataset are shown in Figure 3.4. We used a pre-trained model trained on diverse datasets by Lasinger et al. [203] in our work. A ResNet-based multiscale architecture, as proposed by Xian et al. [206] is used for depth estimation. Adam optimizer is used with a learning rate of 10^{-4} for layers that are randomly initiated and 10^{-5} for layers initialized with pre-trained weights. Decay rates for the optimizer are set at $\beta_1 = .9$ and $\beta_2 = .999$, and training uses a batch size of 8. Due to different image aspect ratios, images are cropped and augmented for training. The input size of the frames is taken as 384×384.

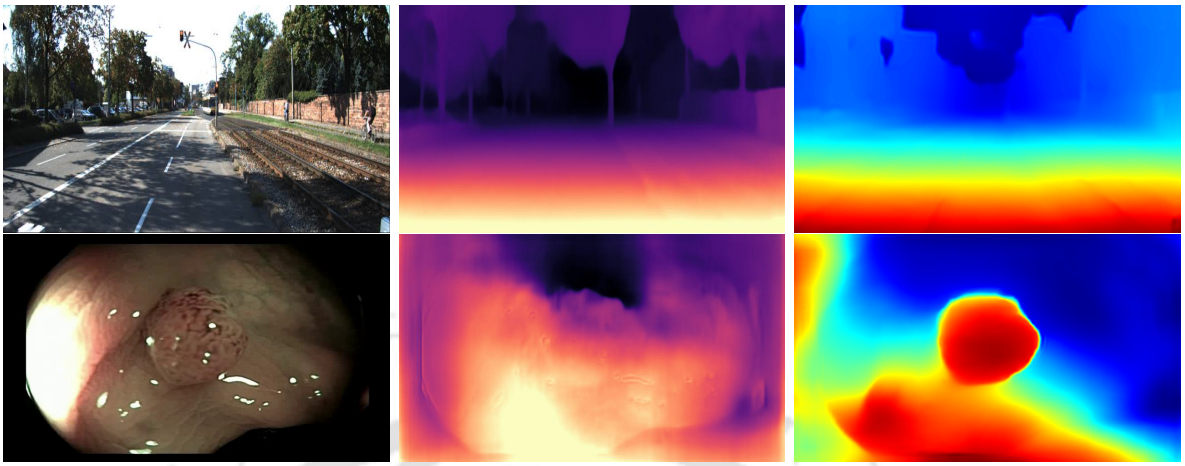


Figure 3.6: Comparison of MDE on two input images, one outdoor and the other one is an endoscopy image. The depth map by Monodepth performs well for outdoor environment while giving unsatisfactory results for the endoscopy image. However, the zero-shot learning method clearly performs well for medical images but cannot accurately estimate the depth in outdoor scenes.

Our method performs better than the state-of-the-art MDE methods. The depth estimation results are shown in Figure 3.6, where the first column represents the input images, while the second and the third column show the comparative results between the monodepth model [199] and zero-shot cross-dataset transfer pre-trained model [203]. This clearly shows that monodepth performs well in outdoor environments than our method. However, the Zero-shot learning method is more accurate in predicting depth in endoscopic images.

Our method is the first of its kind in which key-frames are extracted from an endoscopic video using depth maps. Also, it is robust to occlusions. As redundant frames are discarded in our method, it is more convenient for physicians to analyze essential frames of a video sequence. As explained earlier, the moment distance criterion between consecutive frames is used to ensure that redundant frames are identified and then discarded. The edge magnitude criterion leverages the depth images data to select the best frames. Frames with fewer ORB points have occluded polyps, and these frames are redundant. Adaptive thresholding is used to apply three criteria to obtain essential frames for 3D reconstruction.

The selected key-frames are finally used to reconstruct the 3D surface of the polyp. We have used Facebook's 3D image GUI to view the reconstructed polyp surface; the link to the video is shown here: $https://youtu.be/PJKfk0Mqu2I$. 3D visualization of a polyp helps in surgeries involving the removal of the polyp from its root. This gives better visualization of polyps for diagnosis. Figure 3.5

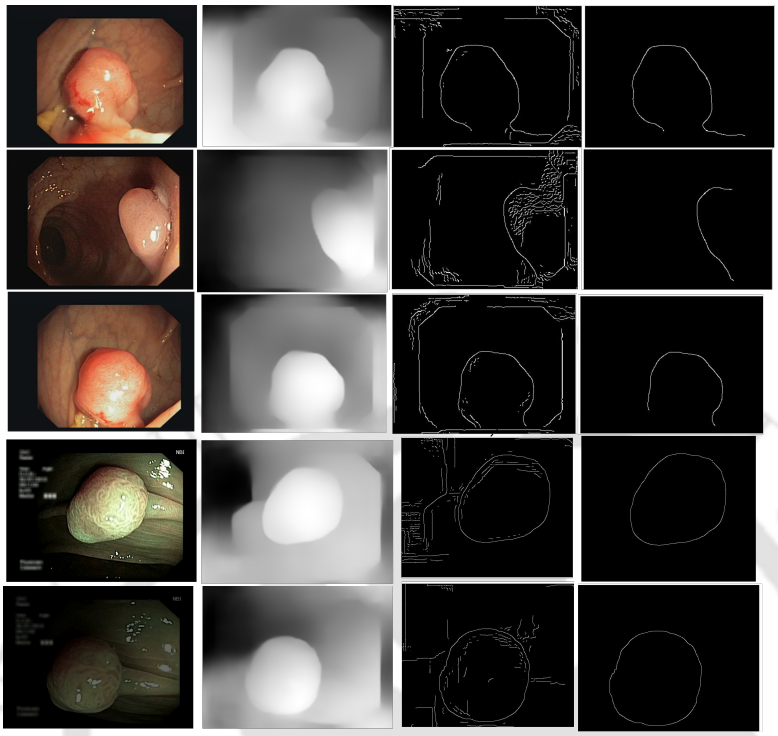


Figure 3.7: Polyp boundary detection using depth map; column 1: original endoscopic image, column 2: generated depth maps, column 3: detected polyp boundary using canny edge detection algorithm, column 4: edge refinement using connected component analysis.

shows some of the results of key-frame extraction and the corresponding depth maps. No publicly available datasets or methods using them that predict depth maps from endoscopic frames exist. Thus, a comparison between different methods for predicting depth from endoscopic images couldn't be performed.

Another application of our proposed method could be the automatic segmentation of polyps in endoscopic images. The depth maps generated by our proposed method can further be used for polyp localization. The canny edge detector is used over the depth maps, and subsequently, the polyp boundary is determined by using connected component analysis. Figure 3.7 shows localized polyps in some of the endoscopic image samples. The segmentation performance on some of the sequences of the CVC-Clinic Database [159] is shown in Table 3.1. This dataset contains 25 colonoscopy video sequences. Each sequence contains an average of 25 frames. We defined mIoU as the mean intersection over the union of the segmented polyp masks to the ground truth masks. In polyp segmentation, an IoU score of ≥ 0.5 is generally considered good [84].

Table 3.1: Key frame selection and segmentation performance using our method on some of the sequences of CVC-Clinic Database (Sequences with only the elevated polyps are considered)

Sequence	#Key frames selected	mIoU
26-50	5	0.501
104-126	7	0.546
127-151	11	0.721
298-317	2	0.723
343-363	7	0.654
384-408	13	0.723
409-428	8	0.663
479-503	20	0.793
504-528	6	0.695
572-591	4	0.698
592-612	5	0.747

3.2.3 Conclusion

Our proposed method can determine depth maps using a zero-shot learning approach. The zero-shot learning method performs well on previously unseen classes like endoscopic images. Through this, we extended MDE to in-vivo images, which would be helpful to analyze medical images. The essential frames are picked out from the colonoscopy videos with the help of depth information and the proposed three criteria selection strategy. The selection of a threshold value for the final fused score must be empirically set to extract the key-frames. Experimental results show the efficacy of the proposed method in selecting key-frames from endoscopic videos and subsequent segmentation of detected polyps in the key-frames with the help of extracted depth maps. Also, the 3D model could be used in clinical diagnosis and surgeries. One possible extension of this work could be the visualization of polyps in detected key-frames in an augmented reality framework.

Though the proposed method can segment the prominent, convex, and elevated polyps that need immediate medical attention, the other polyp structures can not be neglected. They may progress to the malignant stage if not diagnosed early. Therefore, our subsequent studies will focus on segmenting any polyp structures. The following method encapsulates polyps' contextual and spatial information for polyp vs. non-polyp discrimination. The detailed methodology of the technique is discussed in the next proposed approach.

3.3 Adaptive Markov Random Field based Segmentation

Most of the previous works in the domain of polyp segmentation are based on supervised learning. In [207], polyp regions are selected using a Hessian filter and Support vector machine (SVM). Polyp and non-polyp regions are marked based on a threshold λ, which is set experimentally. However, it is very difficult to find an optimum value of λ. Bernal et al. [61] proposed Sector Accumulation-Depth of Valleys Accumulation (SA-DOVA) for polyp segmentation. In [110], a shape-based ultrametric contour map (UCM) is proposed to mark out the polyp regions. However, the method yielded poor segmentation results in some of the cases due to irregular shapes of the polyps. Khan et al. [208] proposed a modified mask Recurrent Convolutional Neural (RCNN) for gastro-intestinal diseases detection like polyp, bleeding, ulcer, etc. However, their model failed for the segmentation of polyps and bleeding regions in most of the cases. Also, their model gives an inferior segmentation performance for the ulcer regions on less training data. In [209], RefineU-Net is proposed for polyp segmentation from colonoscopy images. However, the time complexity of their method cannot be discounted as three deep modules are trained jointly in an end-to-end fashion. All the methods discussed above are supervised and require huge samples for training the models. Also, immense domain knowledge is required to characterize the polyp features. Other deep learning-based techniques used for polyp segmentation are discussed in the result section of this work.

This research work aims at developing an unsupervised method of segmentation using a superpixel-based skeleton. An unsupervised-based method is sometimes recommended over the supervised methods as the latter require large training datasets of manually labeled images, which is sometimes challenging to acquire through an endoscopic procedure. Unsupervised techniques, on the contrary, can be used in the absence of training data for polyp segmentation. Thus, it can be employed during the capturing of colonoscopy videos, i.e., segmentation during the endoscopy. Earlier, we proposed some techniques using the active contour (AC) for unsupervised polyp segmentation [109, 210]. However, these models could provide good segmentation performance for the polyp structure having well-demarcated boundaries. Also, sub-optimal polyp segmentation masks were generated by our methods. Therefore, other polyp features were effectively extracted for their discrimination from the backgrounds in this study.

The proposed segmentation approach can be broadly categorized into two main phases: over-segmenting the image and then aggregating the over-segmented image using a Markov random field

(MRF) framework for final segmentation[2]. The main contribution of this research work is the development of an unsupervised learning framework for polyp segmentation. Firstly, we proposed an adaptive MRF model on graph-based feature interpretation, which is the first work in the domain of colonoscopy image analysis. Superpixel sites are considered the nodes in the region adjacency graph (RAG) to reduce computational complexity. The experimental results show that our method outperforms previous works done on similar datasets. The rest of the work is organized as follows: Section 3.3.1 presents the proposed methodology. Experimental results are given in section 3.3.2. Section 3.3.3 concludes the work.

3.3.1 Proposed method

The proposed adaptive MRF is based on the dominant polyp features. Before describing the implementation of the proposed method, some essential aspects of MRF are discussed below. The proposed adaptive MRF is based on the dominant polyp features.

3.3.1.1 Adaptive MRF

The polyp regions are visually quite different from the rest of the gastrointestinal tract regions. The pixel-level maneuvering is computationally complex, and it yields unfaithful results as the spatial relation among pixels is lost. The proposed superpixel-based method represents a group of similar pixels, and the computational burden is significantly reduced. The overview of the proposed approach is shown in Figure 3.8. Simple Linear Iterative Clustering (SLIC) method [211] is adopted for over-segmenting the polyp frames. To encapsulate the spatial and contextual relation among the adjacent regions resulted after over-segmentation, an Adaptive Markov random field (MRF) frame-work is proposed. Our proposed adaptive MRF is a probabilistic graphical model. It can successfully capture interactions between adjacent regions by encapsulating spatial and contextual similarity information. In literature, MRF is widely used as a tool for semantic segmentation [212–214]. The classical MRF model used for segmentation is pixel-based. However, our proposed method is a superpixel-based model. The details of our proposed model are discussed below:

An essential component of an MRF is defining a neighborhood system. Let Ψ be the set of image lattice sites such that:

$\Psi = \{\psi = (i, j),\ where\ 1 \leq i \leq H, 1 \leq j \leq W \text{ and } i, j, H, W \in\ I\}$, H, W being the height and width

[2]This work has been published in Pattern Recognition Letters, Elsevier, 2022 (Refer *List of publications* page for details).

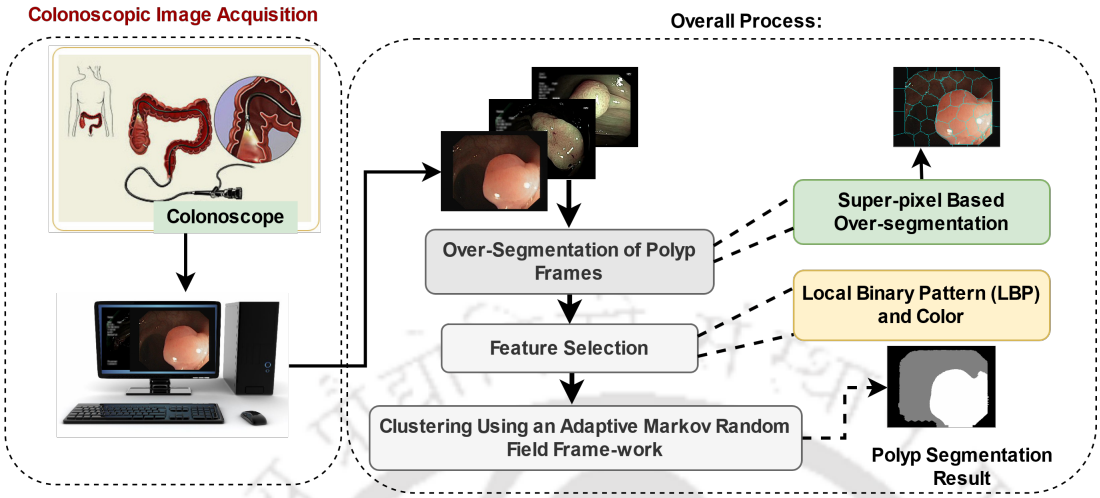


Figure 3.8: Overview of the proposed work; The left side module shows the procedure for colonoscopy image acquisition. The right side block entails different steps involved in our algorithm for polyp segmentation.

(in pixels) of image I, respectively. In the 2-D image lattice Ψ, the pixel values $x = \{x_\psi | \psi \in \Psi\}$ are a realization of random variables $X = \{X_\psi | \psi \in \Psi\}$. A clique ζ is defined as one such subset of Ψ (set of image lattice sites) where every pair of sites are neighbors to each other and a clique set C in the neighborhood system N_ψ is $C = \{\zeta | \zeta \subset N_\psi\}$. Mathematically, a random field X becomes an MRF with respect to the neighborhood system $N = \{N_\psi, \psi \in \Psi\}$ iff:
$P(X = x) > 0 \ \forall x \in \Omega_X$, where Ω_X is the set of all possible x on Ψ; and $P(X_\psi = x_\psi | X_\rho = x_\rho, \ \rho \neq \psi) = P(X_\psi = x_\psi | X_\rho = x_\rho, \ \rho \in N_\psi)$.

According to Hammersley-Clifford theorem, a random field X is a **Gibbs Random Field** (GRF) with respect to the defined neighborhood $N = \{N_\psi, \psi \in \Psi\}$ (where Ψ denotes the set of image lattice sites) iff X is an MRF with respect to the neighborhood. A detailed proof of the same is available in [215]. The theorem allows GRF to globally model local characteristics of an image given by the MRF. A mathematical expression of GRF of a random variable X on the neighborhood system $N = \{N_\psi, \psi \in \Psi\}$ is thus given by:

$$P(X = x) = \frac{1}{Z} \exp\left[-\frac{1}{T} U(x)\right] \tag{3.13}$$

where, $Z = \sum_{x \in \Omega} \exp\left[-\frac{1}{T} U(x)\right]$ is a normalization constant, T is the temperature parameter. $U(x)$ is the energy function with the form $U(x) = \sum_{c \in \Psi} V_c(x)$, where $V_c(x)$ is a potential function.

The segmentation problem can be expressed using the Bayesian framework. Suppose the features

extracted from an image I is denoted as $F = f$, where F is a random variable, and f is an instance of it. Let $Y = y$ be the label field of the segmented image. Then, the problem of segmentation can be posed as the Maximum a posteriori probability (MAP) estimation problem:

$$P(Y = y|F = f) = \frac{P(F = f|Y = y)P(Y = y)}{P(F = f)} \tag{3.14}$$

where, $P(Y = y|F = f)$ is the posteriori probability, $P(F = f|Y = y)P(Y = y)$ is the probability distribution of $F = f$ conditioned over $Y = y$ and $P(Y = y)$ is the prior information.
Suppose feature vector f is of L dimensions, and each component of it is conditionally independent with respect to $Y = y$. Under this assumption, Eq. (3.14) can be re-written as:

$$P\left(Y = y|F = f\right) = \frac{\prod_{l=1}^{L}\left[P(f^l|Y = y)\right]P(Y = y)}{P(F = f)} \tag{3.15}$$

Now, as $P(F = f)$ is known and constant for all the cases, it can be disregarded and the equation can be rewritten as:

$$P(Y = y|F = f) \propto \prod_{l=1}^{L}\Big[P(f^l|Y = y)\Big]\,P(Y = y) \tag{3.16}$$

The first term from RHS of Eq. (3.16) can thus be expressed as i.i.d. MLL or multi-level logistics is used in most MRF-based segmentation models to construct the label distribution. Mostly, the second-order pairwise MLL model is chosen for segmentation. The potentials of all higher-order cliques are set to zeros. Thus, considering that $P(Y = y)$ obeys GRF form, from Eq. (3.13), Eq. (3.16) can again be re-written as:

$$P(Y = y|F = f) \propto \prod_{\psi\in\mathcal{S}} \exp\left[-\Phi(f_\psi|Y = y)\right] \exp\left[\sum_{\ell,\psi\in\mathcal{C}} \theta_{\psi,\ell}(y_\psi, y_\ell)\right] \tag{3.17}$$

where, $\Phi(f_\psi|Y = y)$ is a data penalty term which penalizes a pixel ψ with a label y for given features f. $\theta_{\psi,\ell}(y_\psi, y_\ell)$ is a penalty term used to maintain the smoothness of the label field. It is originally a clique potential function encapsulating the prior probability of labels of the elements of the clique (ψ, ℓ). $N(\psi)$ is the neighborhood of the pixel ψ. Maximizing the expression in Eq. (3.17) is the same as minimizing the following expression.

$$\sum_{\psi\in\mathcal{S}}\Big[-\Phi(f_\psi|Y = y)\Big] + \Big[\sum_{\ell,\psi\in\mathcal{C}} \theta_{\psi,\ell}(y_\psi, y_\ell)\Big] \tag{3.18}$$

The potential function (second term in the above expression) with respect to the labels ($Y = y$) can be modeled as:

$$\theta_{\psi,\ell} = \left[\beta \sum_{\ell,\psi \in \mathcal{C}} \delta(y_\psi, y_\ell)\right] \tag{3.19}$$

where,

$$\begin{aligned} \delta(y_\psi, y_\ell) &= -1 \; if \; y_\psi = y_\ell \\ &= 1 \; if \; y_\psi \neq y_\ell \end{aligned} \tag{3.20}$$

β is a constant, which is specifiable a priori [215, 216].

Unlike the classical MRF, where the neighborhood system is based on pixel sites, our proposed method defines the neighborhood on superpixels. An adjacency graph is built where information about the neighborhood of each superpixel is stored.

Let an image be segmented in such a way that there are q numbers of classes for which the pixels can be grouped. The feature values for each individual class are assumed to follow a Gaussian distribution. After deriving Eq. (3.17) from Eq. (3.16), the term $P(f^l|Y = y)$ for a superpixel s bearing class label m can be expressed as:

$$P(f_s^l|Y = m) = \frac{1}{\sqrt{2\pi}\sigma_m^l} \exp\left(-\frac{(f_s^l - \mu_m^l)^2}{2{\sigma_m^l}^2}\right) \tag{3.21}$$

where, μ_m^l and σ_m^l are mean and standard deviation, respectively, and it is for l^{th} feature component belonging to m^{th} class-label. Taking logarithm of Eq. (3.21) produces a data penalty term, as proposed in Eq. (3.18). The data term contributing to the energy of the features may henceforth be denoted as E_f.

The potential term as depicted in Eq. (3.19) contains β, which is an important parameter. In our proposed work, β is modified to be an adaptive parameter for superpixels. Henceforth, the potential term may be referred to as energy of label, E_l. β is associated with labels in the neighborhood. Attribute variance between neighborhood superpixels is also incorporated to make it practically more plausible. Thus, the proposed weight coefficient β is made adaptive in the following way:

$$\beta^*_{s_\psi, s_\ell} = \frac{\beta_0}{1 + \exp\left(-dist||s_\psi - s_\ell||\frac{\ell}{\psi}\right)} \tag{3.22}$$

where, s_x represent features of x^{th} superpixel, $dist||\cdot||$ is a distance measure, taken as Euclidean distance. So, x denotes the size of the superpixel in terms of the number of pixels within it. β is independent of the input image and is a hyperparameter. The energy minimization functionals have

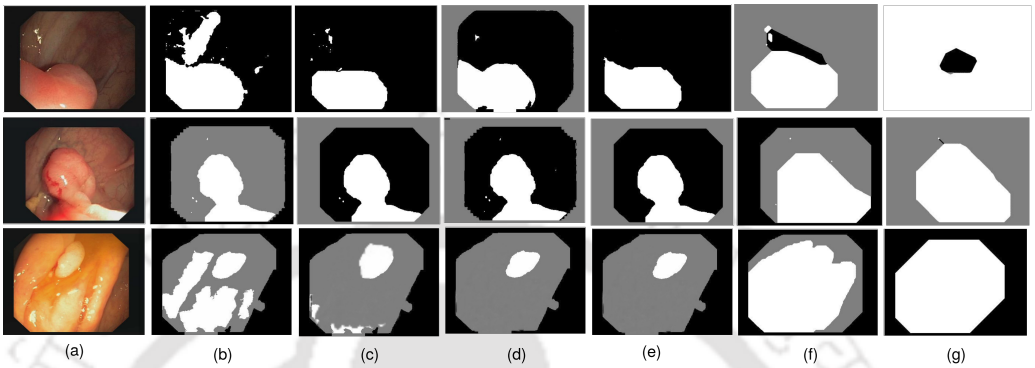


Figure 3.9: Effect of different β values on polyp segmentation results; (a) input image (b) β=1 (c) β=5 (d) β=20 (e) adaptive β value (f) a high β value (g) very high β value.

two components: feature labeling and region labeling components. There are three scenarios which can be considered for the energy components. First, if the constant parameter β makes the region labeling component dominant, the values of estimated parameters μ and σ may deviate much from the feature data, and the segmented result is not consistent. Figure 3.9 (f), (g) demonstrate the outcome using the MLL model alone (which is equivalent to setting β to a very high value) to generate segmentation results. Secondly, if the constant parameter makes the feature modeling component dominant, spatial relationship information would be ignored in the final segmentation result. For example, if $\beta = 0$, the MRF model has the feature modeling component only and cannot produce good segmentation. Finally, both components are considered together by selecting an appropriate value of the constant parameter, and then the best segmentation performance can be achieved. The illustrations of segmentation performances on some samples using different β values are given in Figure 3.9. Figure 3.9 shows that either a small β value or a relatively high β value results in improper segmentation.

Super-pixels belonging to the same class must have a smaller β^*. Similarly, super-pixels belonging to different classes should have more penalty, i.e., a more β^* value. This requirement is satisfied by our proposed adaptive penalty term as given in Eq. (3.22). The initial value of β, i.e., β_0 is set judiciously from extensive experimental results and is set as 20.

3.3.1.2 Implementation of the proposed method

The proposed segmentation algorithm can be treated as an energy minimization problem. The total energy E is given as $E_f + E_l$. Therefore, the whole expression to be minimized is given as:

$$\sum_{s\in\mathcal{S}}\sum_{l=1}^{L}\left(ln\sqrt{(2\pi)^L}\sigma_m + \frac{(f_s^l-\mu_m^l)^2}{2{\sigma_m^l}^2}\right) + \beta^* \sum_{m,n\in\mathcal{C}}\delta(m,n) \tag{3.23}$$

where, $\mathcal{S}$ is the set of all superpixels, f_s^l is l^{th} component of feature vector f_s representing superpixel s, and $\mathcal{C}$ is the set of all cliques carrying doubleton potentials for all superpixel sites. The first term of Eq. (3.23) is represented as E_f, and the second term is denoted as E_l. E_f is the feature modeling component that is used to describe the features of an image. Similarly, E_l is a region labeling component that is used to denote the energy of image regions. The energy of $P(Y = y|F = f)$ is then derived as

$$E_f + E_l \tag{3.24}$$

For the model represented by Eq. (3.23), the MAP will be

$$\begin{aligned}\hat{y} &= \underset{y\in\Omega_Y}{\operatorname{argmax}}\, P(Y=y|F=f)\\ &= \underset{y\in\Omega_Y}{\operatorname{argmax}}\, \frac{1}{Z}\exp\left[-\frac{1}{T}E\right]\\ &= \underset{y\in\Omega_Y}{\operatorname{argmin}}\, E\end{aligned} \tag{3.25}$$

Eq. (13) means that maximizing the posterior conditional probability distribution or Gibbs distribution is equivalent to minimize the model's energy. Therefore, minimizing both the energy terms, E_f and E_l, is vital for optimal labeling or segmentation. If a balance can be achieved between both components by choosing a proper constant parameter, the estimated parameters are usually optimal [217]. Therefore, both the energy terms are essential for final segmentation.

The texture features are encapsulated by employing LBP features. The LBP descriptor is calculated for each pixel inside a particular superpixel, and the mean LBP value is calculated for each of the superpixels. The mean LBP value is also normalized by dividing the mean value by the maximum LBP value of a particular superpixel. The LBP value is calculated on the grayscale image. In our work, a uniform LBP operator is employed. This operator can efficiently extract local image texture information since it is invariant to monotonic grayscale transformations, and also it is easy to implement. Texture features can discriminate polyps and normal regions in colonoscopy frames.

In [44], a uniform LBP operator is employed to extract critical texture features of polyp regions for their detection in wireless capsule endoscopy (WCE) images.

For LBP calculation, $R = 1$ and $N = 8$ are taken, where $R = 1$ is the radius and N is the number of samples considered. In colonoscopy images, polyp regions have higher textural content than normal regions. Similarly, there is a variation in color between the polyp and non-polyp areas [218]. Also, the color of an adenomatous (malignant) polyp is reddish and the color of a hyperplastic (benign) polyps tends to be yellowish [16]. These clinical manifestations shown by the polyps are an inherent characteristic. Therefore, advanced imaging techniques such as the Linked Color Imaging (LCI) modality is used so that the reddish and whitish colors of lesions become redder and whiter, respectively. This can improve the visibility of colorectal polyps, and consequently the polyp detection rates.

The CIE-Lab color space has a high response to these warm colors: red and yellow. Thus, the CIE-Lab color model is used in our work. The superpixels are the nodes in the region adjacency graph (RAG). An initial estimate of the segmentation label field is obtained via k-means clustering. The parameters for energy function are updated via the Expectation maximization (EM) algorithm as given in Eq. (3.26), (3.27), and (3.28). The proposed algorithm is shown in Figure 3.10.

- E-step:

$$P(Y = m|f_s) = \frac{p(f_s|Y = m)p(Y = m)}{\sum_{Y=m} p(f_s|Y = m)p(Y = m)} \tag{3.26}$$

- M-step: parameters are re-estimated

$$\mu^l_{(Y=m)} = \frac{1}{N} \sum_{s,Y=m} f^l_s \tag{3.27}$$

$$\sigma^l_{Y=m}{}^2 = \frac{1}{N-1} \sum_{s,Y=m} (f^l_s - \mu^l_m)^2 \tag{3.28}$$

Optimization in an MRF problem involves finding the maximum of the joint probability over the graph as given by RAG in our method. An efficient technique used for MRF optimization is Iterated Conditional Modes (ICM). It is widely used for segmentation [219]. This algorithm first selects an initial configuration for features, which are represented by the nodes of the graph with labels. As discussed above, the k- means clustering algorithm gives the initial label assignment in our proposed work. Then, it iterates over each node in the graph and tries to reassign new values to minimize the total energy. We have made the convergence adaptive during the energy minimization step using the

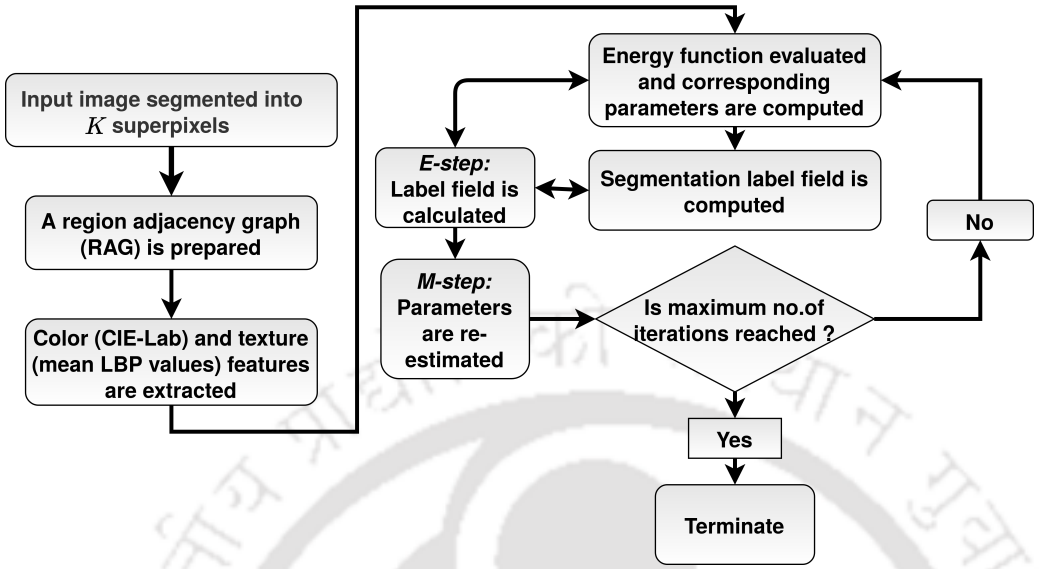


Figure 3.10: Proposed algorithm.

iterative condition mode (ICM). During this operation, the expectation-maximization (EM) step is executed to minimize the total energy, and the segmentation labels are assigned to each superpixel. The E_f and E_l are calculated, and labels are reassigned such that the total energy after each iteration will be minimized. The labels are, therefore, the final outcome of each iteration. Thus, the convergence criterion is based on observing the change in the assigned labels given to the superpixels during ICM. Whenever no change in the labels is noticed for three successive iterations, our algorithm is said to have converged.

3.3.2 Results and discussion

The images used for our experimentation are taken from the CVC-ClinicDB database [159]. The database consists of 612 images, each of size $288 \times 384 \times 3$. It has 29 sequences, each consisting of an average of 25 frames with high intra-frame variations. The database also contains ground truth masks prepared by experts of the Computer Vision Center, Barcelona, Spain. Another database that was also used to validate our proposed method is the ETIS-Larib database [220]. It contains 196 image frames with corresponding manually-annotated ground-truth polyp masks.

The proposed method is tested on each of the sequences of the CVC-ClinicDB dataset. Step-wise results are given in Figure 3.11. In the 2^{nd} column of Figure 3.11, the yellow dots represent the centroids of the super-pixels. Each image is finally segmented into three different classes. The

qualitative polyp segmentation performances on samples of both the databases are shown in Figure 3.12.

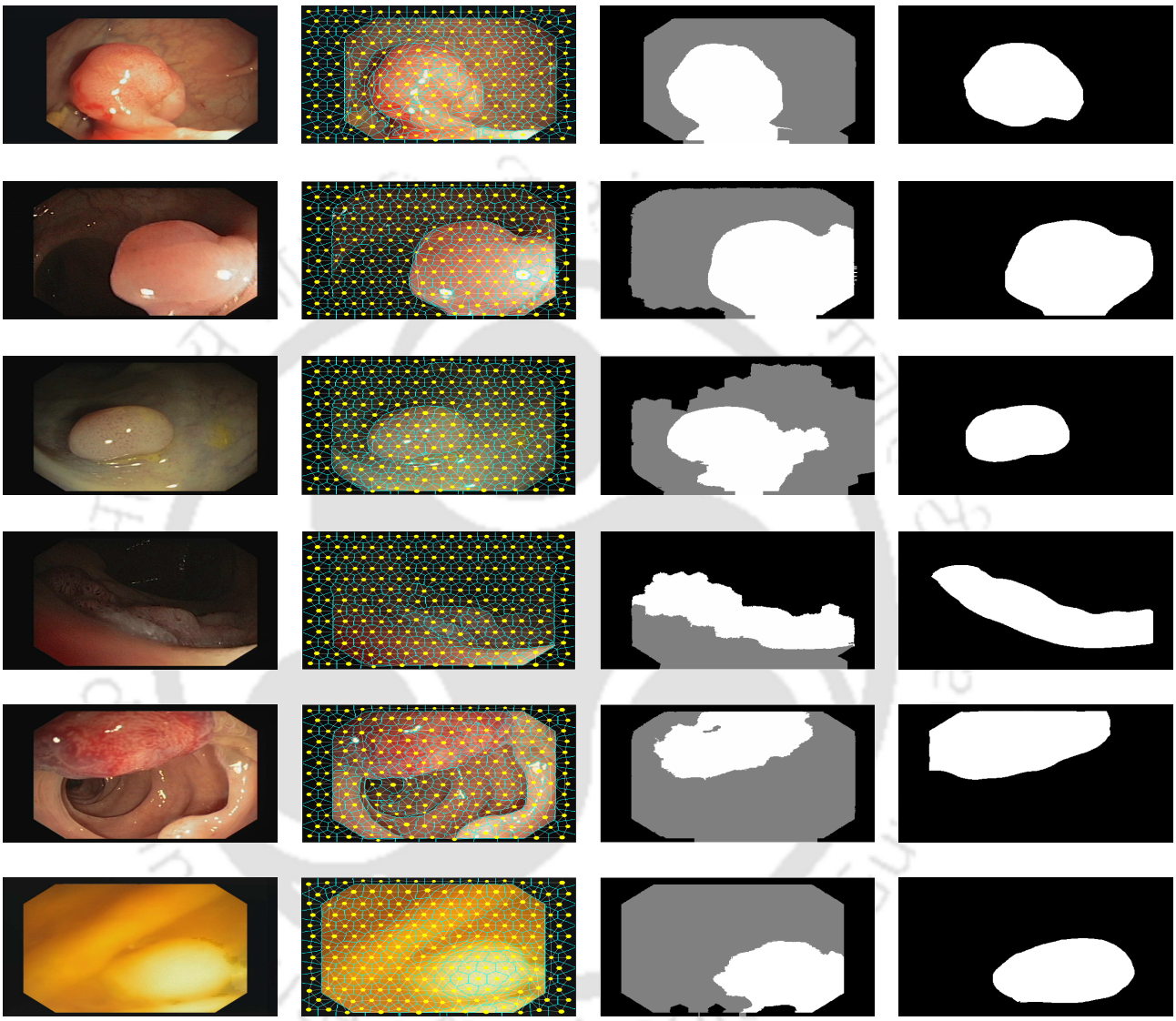


Figure 3.11: In each row, column-wise from left to right: original image; over-segmentation of image with RAG defined on the same; obtained result of 3-class segmentation; ground truth. The proposed method can give satisfactory results in different lighting conditions and camera positions.

Dice Similarity Coefficient (DSC) is chosen to measure the similarity between the final segmentation results and the ground truth. The DSC is expressed in terms of TP, FP, FN as:

$$dice(A, B) = \frac{2 * TP}{2 * TP + FP + FN}$$

where, TP, FP, FN stands for true positive, false positive, and false negative. All the computations

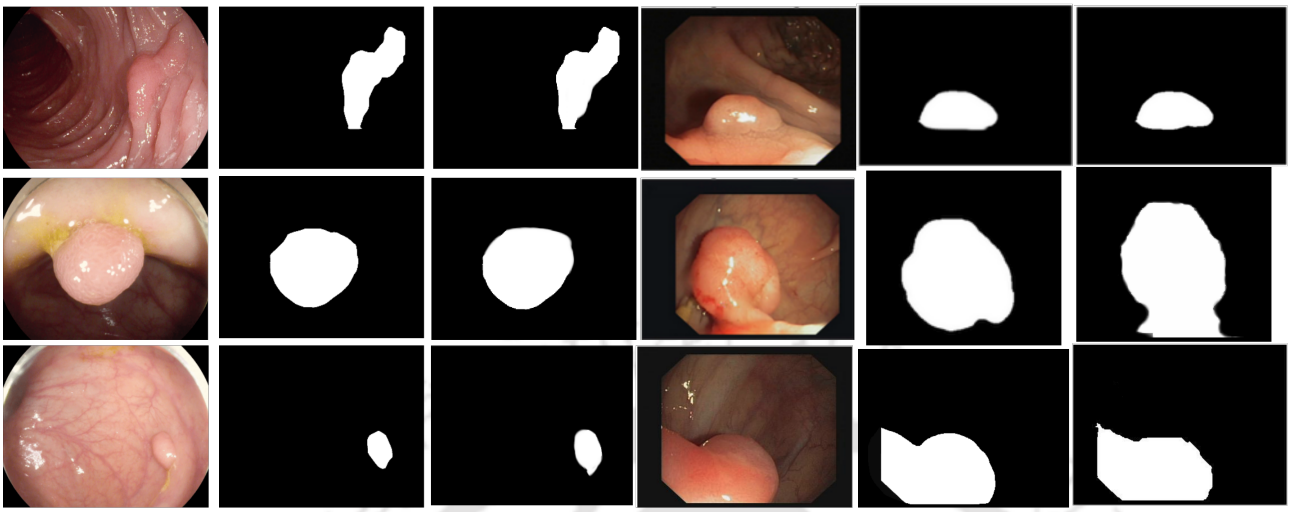


Figure 3.12: Qualitative polyp segmentation performances; (a) samples from ETIS-Larib dataset (b) ground truth masks (c) predicted polyp masks (d) samples from CVC-ClinicDB (e) ground truth masks (f) predicted polyp masks.

relating to accuracy measurement are done at the pixel level. Other performance measures like mean Intersection over Union (mIoU), Recall (Rec.), F1-score (F1), and Hausdorff distance (HD) are also used in this study.

The images of the dataset considered for our study contain very little specularity. Therefore, the effect of specularity is discounted. We also employed specularity inpainting, and the performance was evaluated. We employed a variational inpainting method based on the Mumford–Shah–Euler image model [221]. The specularity masks are generated following the method proposed in [116]. From

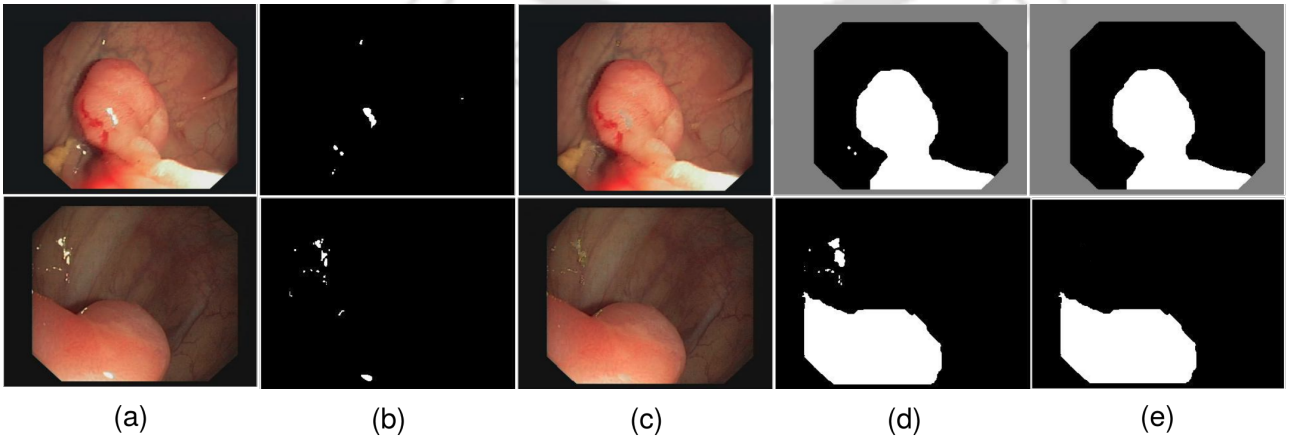


Figure 3.13: Effect of specularity on polyp segmentation; (a) input specular images, (b) generated specular masks, (c) inpainted images, (d) segmentation results before specularity removal, (e) polyp segmentation results after specularity removal.

Table 3.2: Table showing average DSC for different video sequences of the CVC-ClinicDB database.

Seq. No.	Average DSC	Seq. No.	Average DSC
1	0.6841	16	0.5531
2	0.5716	17	0.7601
3	0.4557	18	0.4422
4	0.3457	19	0.7266
5	0.4092	20	0.8181
6	0.5668	21	0.6000
7	0.6234	22	0.5770
8	0.7942	23	0.7812
9	0.4500	24	0.8812
10	0.2156	25	0.6841
11	0.6166	26	0.5593
12	0.5902	27	0.5052
13	0.7036	28	0.8873
14	0.5277	29	0.5947
15	0.7490		

the results, it is seen that the specularity has a nominal effect on polyp segmentation performances. However, the false-positive rate is less for specular free polyp images. Figure 3.13 shows the impact of specularity on the segmentation performances.

For our experimentation, the value of K (number of super-pixels) and k (cluster numbers) were set to be 200 and 3, respectively. A small value of K may not perfectly represent the pixels. Similarly, a very high K value may create unnecessary super-pixels with similar properties. So, a medium-range value for K was considered. The proposed method segments polyp from the non-polyp regions, i.e., two-class segmentation. In some of the colonoscopy frames, black-colored boundary regions are present, which do not convey any pathological information. The parameter k controls the extent of image segmentation. For $k = 2$, frames are under segmented, whereas for $k > 3$, the polyp frames are over segmented, as shown in Figure 3.14. The pixel-based MRF with a similar set up takes 14.83 seconds for processing a colonoscopy frame. The average processing time in seconds for processing a 288×384×3 image frame for generating superpixel-based segmentation is 5.3435 sec/frame by our algorithm. Similarly, an additional 3.2010 sec/frame is needed for final segmentation. Overall, a total of 8.53435 sec is required to get the final segmentation output. The experiments were simulated using Matlab 2017 with an Intel i3, 1.80GHz processor, and a 4 GB RAM system. Our method is an

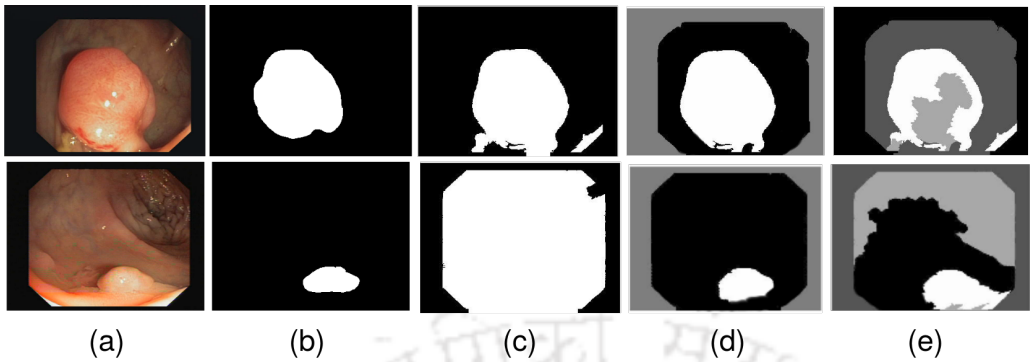


Figure 3.14: Selection of number of classes for polyp segmentation. (a) input image (b) ground truth (c) $k = 2$ (d) $k = 3$ (e) $k = 4$

unsupervised learning approach, so no prior learning is required. Our proposed system can process real-time colonoscopy frames and provide a segmentation score for each polyp frame in a GPU-based system. The proposed method achieves a mean DSC of 60.77%, which is reasonably high compared to state-of-the-art techniques. Table 3.2 shows the segmentation performance in terms of the average DSC value for each of the 29 sequences of the CVC-ClinicDB. Our method was also validated by another publicly available dataset, which is ETISLarib polyp DB [220]. Our approach gives an mIoU of 41.48% in polyp segmentation of 196 colonoscopy frames.

Table 3.3: Comparision with the state-of-the-art methods

Methods	**DSC**	**mIoU**
MSA-DOVA [176]	36.27	22.13
SA-DOVA [61]	55.33	37.93
FCN8 (2 class) [77]	67.44	50.85
Shape-UCM [161] (87 polyps)	65.77	49.0
GPB-OWT-UCM [177] (87 polyps)	61.11	44.0
Proposed (612 polyps)	60.77	42.85

Deep learning-based methods are now getting attention in medical image analysis. Segmentation of endoscopic images using FCN8 architecture [77] gives an IoU of 50.85% for polyp segmentation. The UNet [175], which is used extensively for medical image segmentation, gives a mIoU of 0.4711 in polyp segmentation for the CVC-ClinicDB database. Similarly, ResUNet [222] gives a mIoU of 0.4570 for the same dataset. Table 3.4 shows the segmentation performance of some of the state-of-the-art deep baseline models on the CVC-ClinicDB database.

Deep learning-based methods generally give better accuracy. However, most of the deep learning-

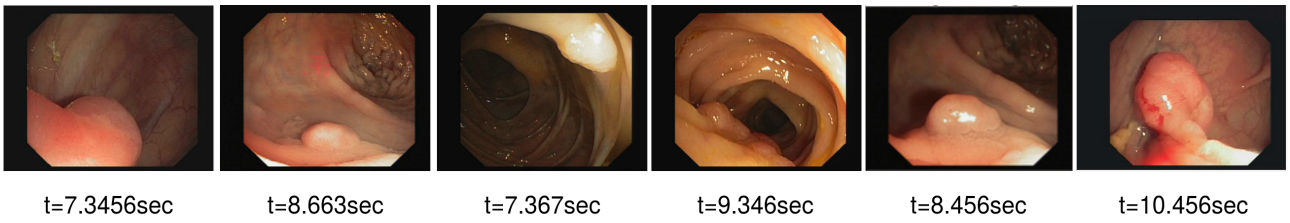


Figure 3.15: Time complexity of the proposed algorithm on different polyp structures.

based methods reported in the literature have trained and tested the models with a common set of data. Also, these models require a considerable number of training samples and ground truth masks during the model training. In some instances, the ground truth information may not be readily available. Also, the image features may change significantly across different imaging modalities, which needs retraining of the already trained models. In contrast, we proposed an unsupervised polyp segmentation which can give a reasonable performance even for unseen data. Though the DSC of the proposed method is a little inferior to most state-of-the-art techniques, the F1 is competitive. In polyp segmentation, it is to be mentioned that accurate delineation of polyp is not required for most pathological interpretations. Our approach gives a higher recall (Rec.) value. Table 3.3 and Table 3.4 show the performances of some of the state-of-the-art methods on CVC-ClinicDB. Our proposed method achieves a competitive IoU value with respect to the deep learning-based approaches. Figure 3.15 shows some of the image

Table 3.4: Comparision of segmentation performance with different baseline models and state-of-the-art methods on CVC-Clinic Database

Method	DSC	Rec.	F1
U-Net [175]	0.756	0.659	0.721
DeepLabv4 [223]	0.8434	0.759	0.794
PraNet [224]	0.7256	0.629	0.756
cGAN [225]	0.8235	0.732	0.789
ResUNet [222]	0.451	0.612	0.654
LGWe-LSM [116]	0.754	0.725	0.750
Hybrid-CNN [114]	0.734	-	-
Proposed	0.6077	0.761	0.732

samples of CVC-ClinicDB and the processing time for polyp segmentation.

To provide a fair comparison of our method, we compare our results with some of the state-of-the-art segmentation results. The results given by Shape-UCM and GPB-OWT-UCM methods in Table 3.3 show the segmentation performance on some NBI image datasets [161]. In these methods, the polyps are manually placed in the center of the frame, which may not be applicable during real-time

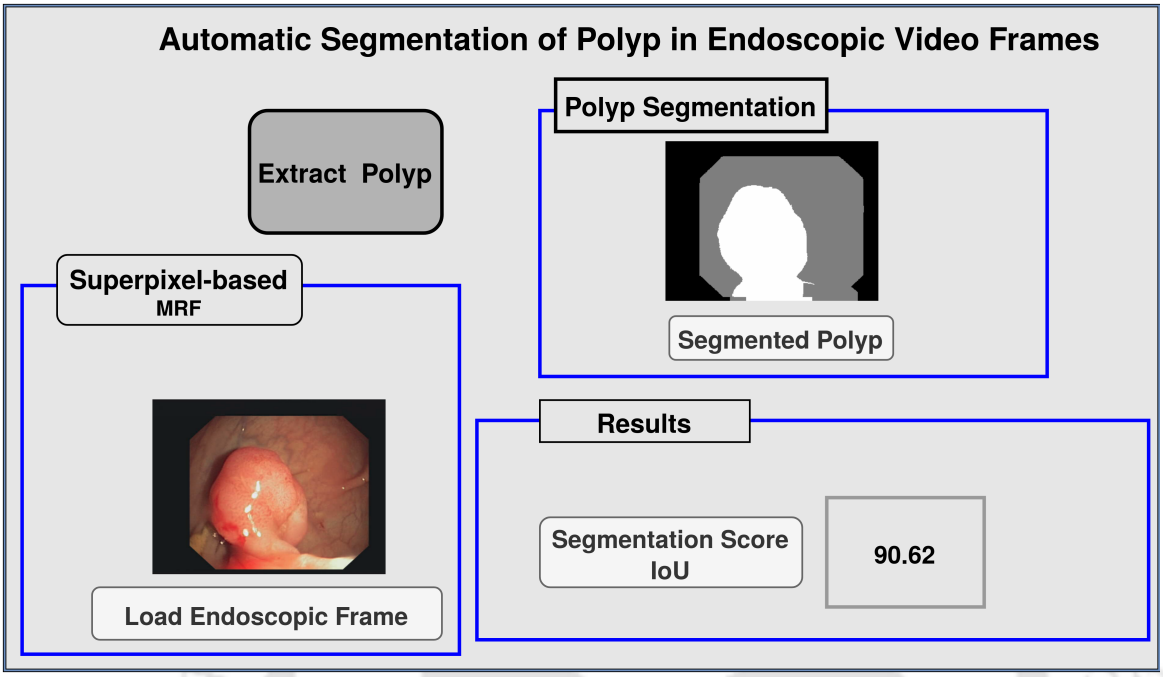


Figure 3.16: The designed GUI for polyp segmentation.

analysis. All the state-of-the-art methods given in Table 3.3 are supervised, whereas our proposed method is unsupervised. Also, most methods are tested on a very small dataset that is not publicly available. Methods deployed in [61, 77, 176] used CVC-ClinicDB dataset for segmentation.

We employed some of the recent deep models for polyp segmentation in our study. U-Net, DeepLab, PraNet, Polypnet are some of the popular deep architectures, which have been employed for polyp segmentation. The CVC-ClinicDB was split, 90% was used for training, and the rest 10% was used to test these models. The DeepLabv4 gives a DSC of 0.8434. The U-Net gives a segmentation score (DSC) of 0.7561, and the PraNet provides a DSC of 0.7256 on the CVC-ClinicDB. Figure 3.17 shows the qualitative polyp segmentation performances of these deep models. Table 3.4 shows the quantitative measures of some of these deep models on polyp segmentation.

Table 3.5: Comparision of segmentation performance in terms of Hausdroff distance (HD) with different baseline models and state-of-the-art models on CVC-ClinicDB Database.

Method	**HD**
DT-WpCNN [116]	112.615
LGWe-LSM [116]	32.848
U-Net [175]	43.254
ResNet50 [142]	55.198
Polypnet [116]	21.796
Proposed	48.55

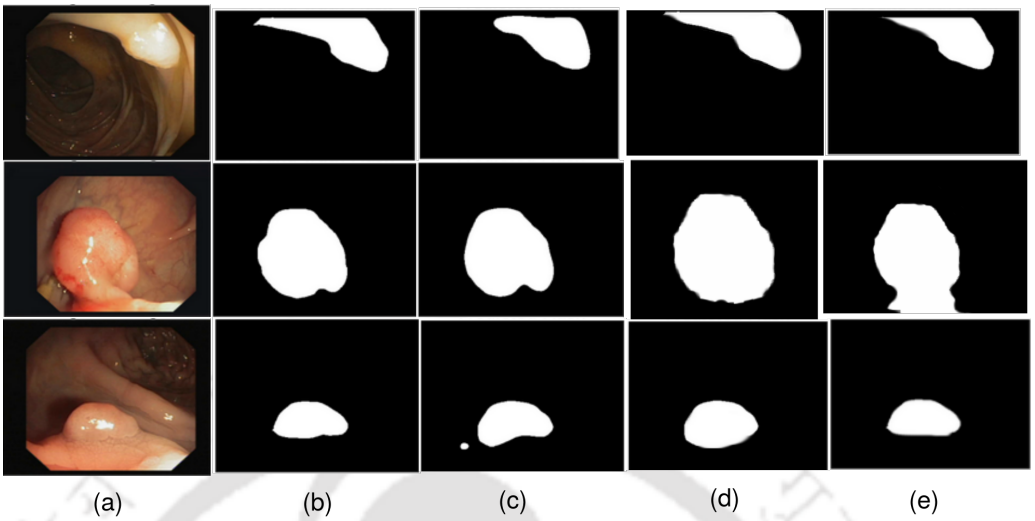


Figure 3.17: Polyp segmentation performances; (a) input images, (b) ground truth masks, (c) DeepLabv4, (d) PraNet, (e) Adaptive MRF.

Figure 3.18: Hausdorff distance measure between the ground truths and the respective predicted polyp masks.

Hausdorff distance (HD) measures the distance between the ground truth surface and the segmented surface. A smaller HD indicates better segmentation accuracy. The average HD using the proposed method is 48.5532. The HD measures on the segmented output of two polyp samples are shown in Figure 3.18. The HD values for some of the recent polyp segmentation methods on CVC-ClinicDB are compared with the HD of the proposed method in Table 3.5. A GUI based on the proposed algorithm is designed as shown in Figure 3.16.

3.3.3 Conclusion

The proposed method can produce better segmentation performance with a reduced computational load than the pixel-based method. The MRF parameter β has been made adaptive based on the

dominant image characteristics in our approach. LBP and Lab color features are incorporated into the MRF model. The proposed endoscopic polyp segmentation method is unsupervised, whereas most of the state-of-the-art techniques are supervised. Deep learning-based methods need substantial datasets for training the networks. However, our proposed unsupervised method can perform well even in a small dataset, which may be suitable for endoscopic image analysis in real-time. Our algorithm sometimes misses partially occluded polyps and polyps that are texturally indistinctive to the complex background. Similarly, polyps that are very small in size may result in an over-segmentation outcome by our method. Therefore, an approach using context and frame correlation information in a colonoscopy video sequence can be employed for polyp segmentation in the future. This could better segment polyps and can handle partial occlusion. The convolutional neural network (CNN) could efficiently extract global and local contextual information. Other geometrical information could be integrated for polyp segmentation, like polyp shape information.

Our previous proposed approaches have studied the utilities of saliency map in polyp detection and segmentation. In the subsequent chapter, we will explore the effect of polyp shape information on polyp boundaries detection. Based on these propositions, a saliency map-guided shape compactness is formulated for polyp segmentation in the next chapter. As discussed earlier, the polyp is the salient region in the colonoscopy frame. The polyps (ROIs) in the frame are firstly highlighted in the saliency or probability maps. A deep learning framework is employed to generate the saliency map for polyps as it can extract generic features from various polyp structures. Shape information of the polyp is further utilized for better segmentation.

3.4 Saliency Map-Guided Shape Compactness for Segmentation

A detection strategy using a saliency-based selection of candidate polyp regions was proposed by Deeba et al. [226]. They tried to enhance the saliency of clinically significant features in endoscopic images for polyp detection. On laryngeal endoscopic images, Ding et al. [227] suggested a deep attention network built on U-Net with color normalization operation (CN-DA-UNet) for generation of probability map followed by thresholding for glottis segmentation. Shape descriptors can be used as the features for semantic segmentation in an image [228,229]. The shape and appearance of lesions are essential for the early diagnosis of polyp in the GI tract [38,62]. Shape compactness (SC) is one such shape descriptor used widely for medical image segmentation [230].

The polyp regions are texturally and visually different from other regions. An efficient approach for generating the saliency map is crucial for better polyp segmentation. Therefore, a U-Net architecture is employed in the current work to generate a probability map. Upon the generated probability map, a shape constraint prior is imposed. The resultant combination is formulated as an energy minimization problem. This problem is solved using the alternating direction method of multipliers (ADMM) for final polyp segmentation[3]. Section 3.4.1 discusses the proposed methodology. Experimental results and conclusions are elucidated in section 3.4.2 and section 3.4.3, respectively.

3.4.1 Proposed method

The saliency map-based methods are effective for object detection. It provides probable locations of the target objects in an image by highlighting the ROIs. Therefore, locating polyps in a colonoscopy image using a saliency map seems to be effective. The overall method is shown in Figure 3.19. The details of the method are discussed below.

3.4.1.1 Probability map generation

The polyps are the salient regions in the endoscopic frames. U-net deep architecture is one of the models used extensively for biomedical image segmentation [175]. U-Net is considered efficient in object segmentation, even in limited training data scenarios. Therefore, in this current work, we propose to employ the U-Net-based deep convolutional network to generate probability maps as a unary potential function for segmentation. This method produces promising segmentation results,

[3]This work has been published in Signal, Image and Video Processing, Springer, 2022 (Refer *List of publications* page for details).

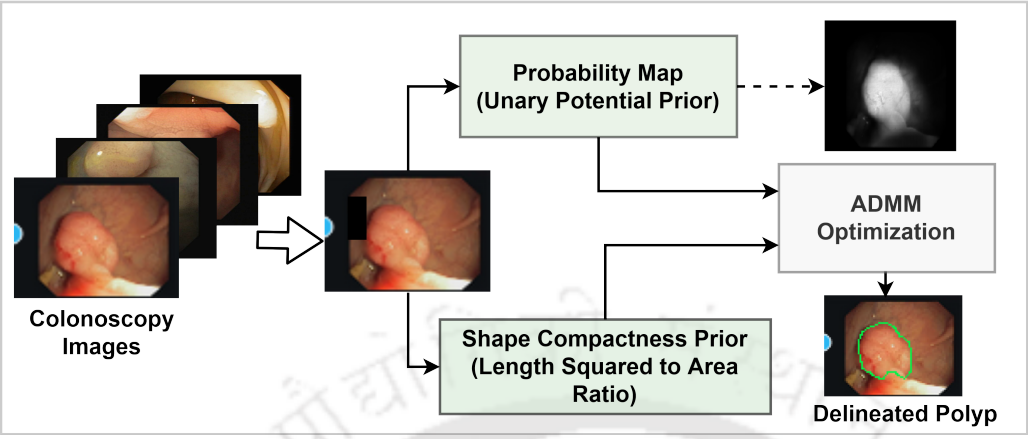


Figure 3.19: The proposed pipeline of polyp delineation in colonoscopy video frames.

which uses skip architecture [231]. The generation of the probability map for segmentation has been inspired by the work of Ronneberger et al. [175].

3.4.1.2 Geometric shape compactness prior

A common shape compactness is usually represented as: $S = \frac{(\text{Perimeter})^2}{\text{Area}}$, where S is the measure of compactness or roundness. The measures of shape compactness are invariant under translation, rotation, and scaling transformations, dimensionless, and minimized by a circle [229]. Our current work is inspired by the work of Montero et al. [229].

In the case of endoscopic polyp segmentation, we must use a dimensionless, unbiased, and position-independent shape compactness prior to constrain the segmentation functionals. The polyp shape is mostly circular; the size varies in a wide range. In our proposed method, the ratio of *length-squared* to the area is used as the compactness prior. This prior is a widely used shape compactness measure in shape metrics [229, 232]. Such functionals pose an NP-hard optimization problem, which can easily be solved by an alternating direction method of multipliers (ADMM). The detail of the method is given in section 3.4.1.3.

3.4.1.3 Implementation

Let Λ represent a 2-Dimensional image domain and $\mathbf{m}_j \in \mathbb{C}^R$ be the input feature vector of pixel $j \in \Lambda$. Segmentation assigns each $j \in \Lambda$ a label $\mathbf{x}_j \in L$. Only two labels are considered in our method, i.e., 1 for polyp and 0 for non-polyp regions. The optimal labeling can be formulated into an energy

minimization problem. Mathematically, it is represented as :

$$E(\mathbf{x}) = E_{up}(\mathbf{x}) + \lambda E_{cs}(\mathbf{x}) \tag{3.29}$$

Where, binary vector $\mathbf{x} \in \{0, 1\}$, E_{up} is the energy term that minimizes entropy and E_{cs} is the energy imposed by a shape prior. E_{up} is contributed by each pixel attribute, called unary potential prior.

In this work, a detailed study is carried out on unary-potentials. The unary-potential term is given by a probability map. The output of a probability map indicates the probability of each pixel belonging to a particular label subjected to some conditions. The other energy prior, i.e., E_{cs} is the shape compactness prior, which is the ratio of *length-squared* to the area, i.e.,

$$E_{cs}(\mathbf{x}) = P(\mathbf{x})^2 / A(\mathbf{x}) \tag{3.30}$$

In digital domain, the length can be expressed in terms of the number of neighbouring pixels with distinct labels, and the area may be represented as $A(\mathbf{x}) = \sum_j x_j$ i.e.,

$$P(\mathbf{x}) \propto \sum_{j,i} \beta_{ji}(x_j - x_i)^2 \tag{3.31}$$

Where, the pairwise clique potential term is defined as:

$$\beta(ji) = \begin{cases} 1, & \text{if } j, i \text{ are neighbors.} \\ 0, & \text{otherwise} \end{cases}$$

This weighing function can be different according to the context. The compactness measure is formulated to converge the solution towards strong edges in the image. The weight matrix is given by :

$$\beta_{ji} = \exp\left(-\sum_k \sigma_k \left(x_{jk} - x_{ik}\right)\right)^2 \tag{3.32}$$

Where, σ_k controls the relative importance of feature k on the weight. The length is expressed as $P(\mathbf{x}) = \mathbf{x}^\top L_m \mathbf{x}$, where L_m is the Laplacian matrix corresponding to weight β_{ji}. Thus, the compactness model takes the final form as :

$$\underset{x \in \{0,1\}}{\arg\min} \;\; E(\mathbf{x}) = \mathbf{u}^\top \mathbf{x} + \lambda \frac{\left(\mathbf{x}^\top L_m \mathbf{x}\right)^2}{\mathbf{1}^\top \mathbf{x}} \tag{3.33}$$

Where, $\mathbf{1}$ is a vector with value one for each element. The energy functional given in Eq. (3.33) is a non-convex optimization problem and is very difficult to get the closed form solutions. The

ADMM is an approach for solving such problems by breaking them down into smaller parts that are relatively easier to handle. It introduces equality constraints and makes it a convex form via augmented Lagrangian [233]. Therefore, with the introduction of variable $\mathbf{z} \in C^{\Lambda}$ and $t \in C_{+}$, Eq. (3.33), which was an in-equality problem is now can be written as:

$$\begin{aligned} \underset{\mathbf{x},\mathbf{z},t}{\arg\min} \;\; & \mathbf{u}^{\top}\mathbf{x} + \frac{\lambda}{t}(\mathbf{x}^{\top}L_m\mathbf{x})(\mathbf{z}^{\top}L_m\mathbf{z}), \text{ s.t } \;\; \mathbf{x} = \mathbf{z}, \\ & \text{and } \;\; t = \mathbf{1}^{\top}\mathbf{z} \end{aligned} \tag{3.34}$$

Thus, Eq. (3.34) can be solved via introduction of augmented Lagrange formulation:

$$\underset{\mathbf{x},\mathbf{z},t,\gamma_1,\gamma_2}{\arg\min} \;\; \mathbf{u}^{\top}\mathbf{x} + \frac{\lambda}{t}(\mathbf{x}^{\top}L_m\mathbf{x})(\mathbf{z}^{\top}L_m\mathbf{z})) + \frac{\mu_1}{2}||\mathbf{x} - \mathbf{z} + \gamma_1||_2^2 + \frac{\mu_2}{2}||s - 1^{\top}\mathbf{z} + \gamma_2||_2^2 \tag{3.35}$$

Where, γ_1 and γ_2 are dual variables and μ_1, μ_2 are controlling parameters. This formulation allows solving each variable while considering other variables as constant. An iterative optimization method is adopted to solve this problem:

(i) *Updating* $\mathbf{z}$: a sparse-matrix inversion based on Woodbury identity, solved via the preconditioned conjugate gradients method [232].

(ii) *Updating* t: t is given by a closed-form solution of a cubic equation, and

(iii) *Updating* $\mathbf{x}$: $\mathbf{x}$ is updated by solving a graph-cut problem using Boykov-Kolmogorov algorithm [234]. The detailed mathematical formulation can be found in [232].

3.4.2 Results and discussion

3.4.2.1 Datasets

The proposed method is evaluated on two publicly available datasets.

CVC-ClinicDB: It contains 612 challenging images with ground truth, taken from 29 video sequences [159]. Some of the images from the dataset are shown in Figure 3.20. Visually different polyp characteristics are seen in this database.

ETIS-Larib Polyp DB: It contains colonoscopy frames of size 1225×966 along with the corresponding ground-truth masks. A total of 196 frames are available with this dataset [159].

For training the U-Net-based deep model, the following dataset is used.

CVC-ColonDB: This dataset contains 300 images of size 574×500 with pixel-level annotated polyp masks for each image [61].

Some of the example samples from the databases are shown in Figure 3.20.

Table 3.6: Segmentation performances comparision between adaptively thresholded saliency maps generated using state-of-the-art methods and our proposed framework on both the datasets.

Methods	CVC-ClinicDB				ETIS-Larib Polyp DB			
	DSC	Precision	Recall	F1-score	DSC	Precision	Recall	F1-score
SDSP [172]	0.47	0.45	0.54	0.49	0.29	0.38	0.52	0.43
HDCT [235]	0.45	0.39	0.63	0.48	0.27	0.29	0.40	0.33
GBVS [236]	0.24	0.27	0.34	0.30	0.21	0.24	0.35	0.28
Saliency Map [237]	0.66	0.78	0.74	0.75	0.62	0.72	0.79	0.74
SDSP+SC	0.53	0.43	0.67	0.52	0.41	0.49	0.51	0.50
HDCT+SC	0.49	0.45	0.65	0.53	0.34	0.51	0.44	0.47
GBVS+SC	0.43	0.34	0.52	0.41	0.30	0.31	0.47	0.44
Saliency Map+SC	0.71	0.75	0.78	0.77	0.63	0.61	0.79	0.68

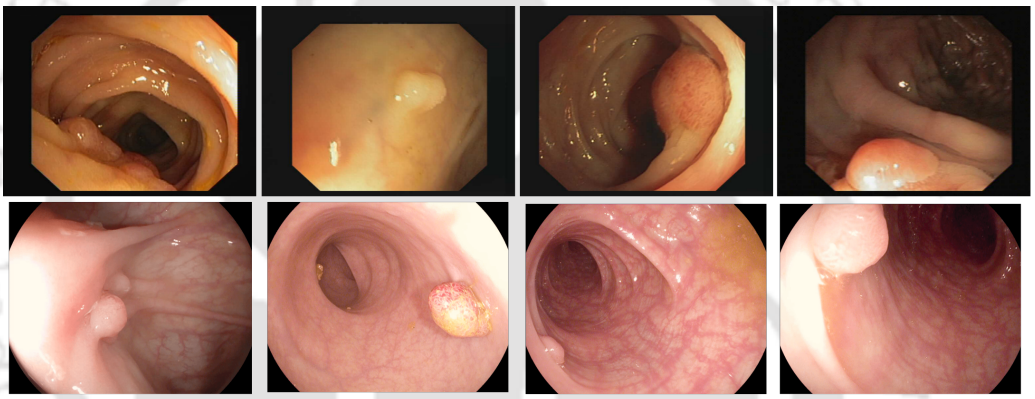


Figure 3.20: Some of the image samples; top row images are from CVC-ClinicDB and bottom row images are from ETIS-Larib Polyp DB.

Our method provides a competitive segmentation performance compared to the existing methods. Our approach uses the polyp shape in compactness prior, and the saliency map generated using global features in a colonoscopy frame. Therefore, both geometrical and textural information is utilized in the proposed approach. Polyps have markedly different colors and textures compared to the non-polyp regions. The U-Net extracts this information to generate probable salient regions for the polyps. The saliency map gives approximate polyp boundary information. Using the shape of the polyp as prior information to the already generated saliency map, we could able to delineate the polyp boundaries perfectly. The U-Net is very efficient in medical image segmentation. The efficiency of the segmentation is further enhanced with the introduction of shape compactness prior to the framework.

In our work, the segmentation of endoscopic polyps is done by formulating an energy minimization problem. The energy terms as in Eq. (3.29) are solved by the ADMM. The equation consists of a unary potential term and shape constraint term. The probability map which accentuates the polyp regions

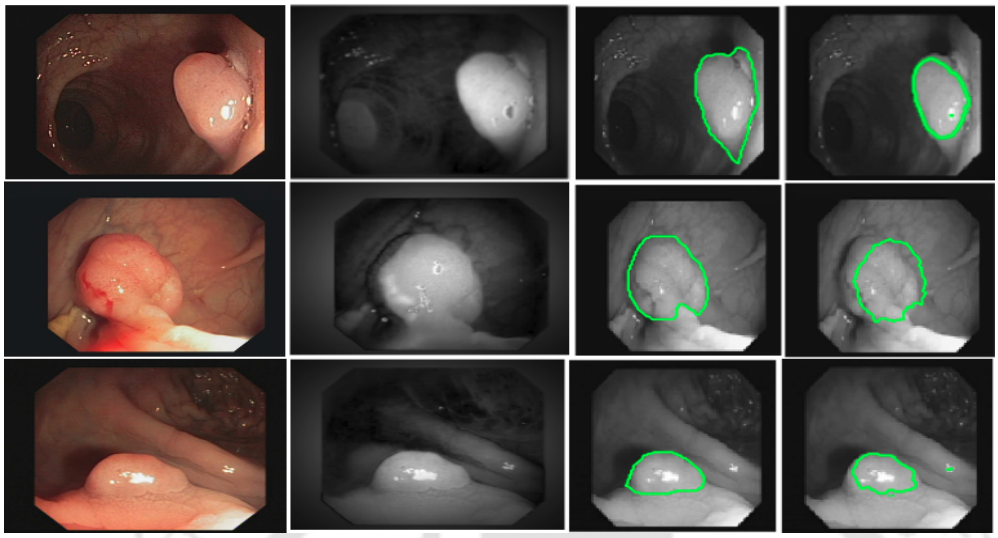


Figure 3.21: Polyp segmentation results on some image samples; from left to right –1^{st} column: Colonoscopy frame, 2^{nd} column: Saliency map using SDSP, 3^{rd} column: Ground truth, and 4^{th} column: ADMM segmentation results.

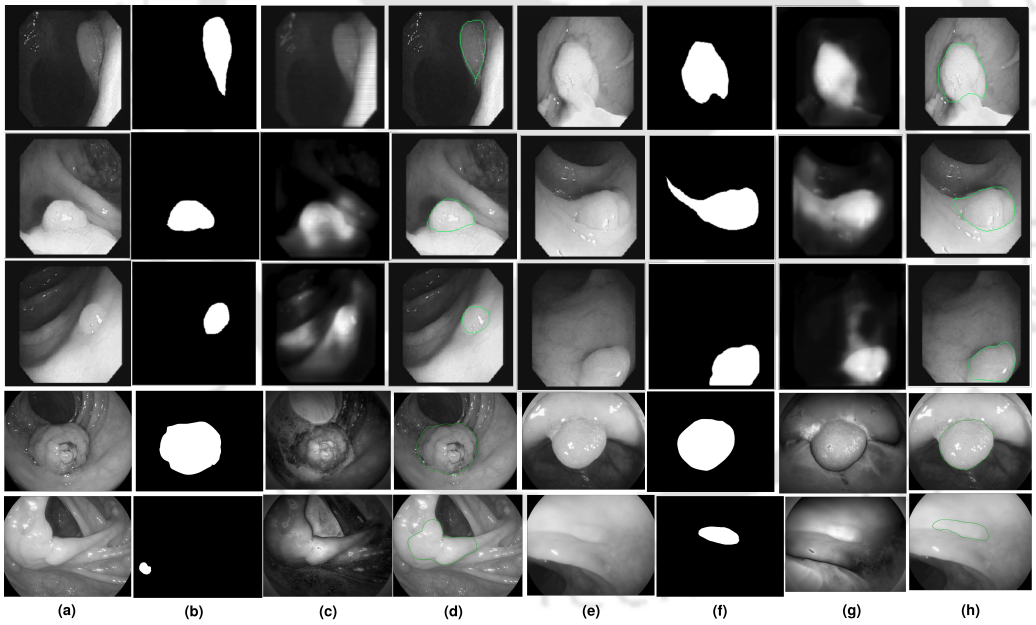


Figure 3.22: Qualitative performances of polyp segmentation: (a), and (e) Colonoscopy frames; (b), and (f) Corresponding ground truth masks; (c), and (g) Saliency maps generated using U-Net; (d), and (h) Final segmentation results using SC.

represents the unary-potential term. Figure 3.21 shows the qualitative results of segmentation using our proposed algorithm on some image samples. The saliency maps shown in this figure are generated using a state-of-the-art SDSP method [172]. Any model that generates the probability map can be used as the unary potential term in the objective function as given in Eq. (3.29). A better saliency

Table 3.7: Comparision of segmentation performance with different baseline models and other individual models on CVC-Clinic Database

Method	DSC	Pre	Rec	F1
MSA-DOVA [176]	0.362	-	-	-
SA-DOVA [159]	0.553	-	-	-
FCN8 (2 class) [77]	0.674	-	-	-
Shape-UCM [161]	0.657	-	-	-
GPB-OWT-UCM [177]	0.611	-	-	-
Saliency Map [237]	0.660	-	-	-
DT-WpCNN [116]	0.809	0.815	0.728	0.769
LGWe-LSM [116]	0.754	0.778	0.725	0.750
FCN (32 s) [238]	0.643	-	-	-
FCN (16 s) [238]	0.689	-	-	-
CPFNet [239]	0.801	-	-	-
U-Net [175]	0.767	0.798	0.659	0.721
ResNet50 [240]	0.718	0.704	0.612	0.654
Hybrid-CNN [114]	0.834	-	-	-
Polypnet [116]	0.839	0.836	0.811	0.823
Proposed	0.821	0.801	0.721	0.759

Table 3.8: Comparision of segmentation performance with different baseline models and other individual models on ETIS-Larib Database

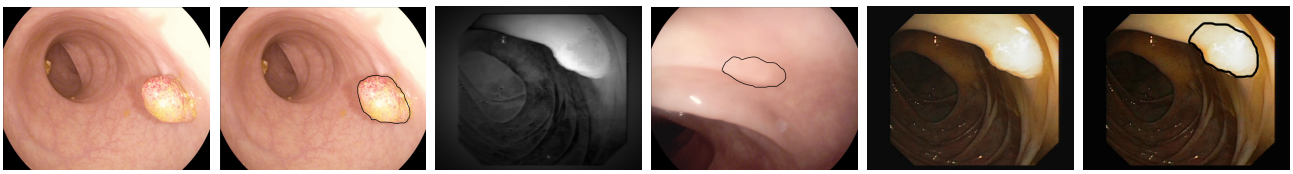
Method	DSC	Pre.	Rec.	F1
Saliency Map [237]	0.6301	0.685	0.702	0.6929
FCN (32 s) [238]	0.610	-	-	-
FCN (16 s) [238]	0.645	-	-	-
CPFNet [239]	0.765	-	-	-
U-Net [175]	0.702	0.756	0.689	0.720
ResNet50 [240]	0.708	0.687	0.645	0.665
Hybrid-CNN [114]	0.812	-	-	-
Proposed	0.756	0.786	0.678	0.728

map would produce a better final segmentation score, which can also be validated from the quantitive and qualitative results. Our method employed the U-Net architecture to generate the probability map as the unary potential energy. The detail of the network architecture is given in [175]. The other energy term is given by $length - squared$ to area shape constrained prior. The energy is minimized by ADMM optimization. The parameters used in the experimentations are: $\mu_1 = 200$, and $\mu_2 = 50$. The compactness regularizer λ was set at 5000. For calculation of shape compactness prior, σ=25 was taken. As given in [241], the ADMM parameters are very stable and do not change the results with the change of these parameter values. A detailed study on finding segmentation results using different probability maps is done in our work.

HDCT [235], and GBVS [236] which are some of the efficient state-of-the-art methods for the

generation of probability maps are considered for comparative analysis. After the generation of saliency maps, otsu's method is used to find the optimal threshold value of segmentation for these methods [235] [242]. The segmentation performance of all these methods is given in Table 3.6. To further improve the segmentation performances, the shape compactness prior was added and formulated using the ADMM. The improvement in results is reported in Table 3.6.

To justify the impact of our proposed framework on each of the approaches, we compared the segmentation performance of our model to each of individual methods as well as certain baseline CNN's often used in the medical image analysis, namely U-Net [175] and ResNet-50 [240]. Each of the metrics quantifies and assesses the segmentation method's effectiveness. For each of the tabulated methods, Table 3.7 shows the mean DSC, Precision, Recall, and F1-score. Our technique outperforms most of the existing methods in terms of different parameters. Overall, the performance of our suggested approach is comparable to that of state-of-the-art methods. The CVC-ColonDB dataset is used for training, whereas the CVC-ClinicDB is used for testing. It is to be mentioned that the CVC-ClinicDB database is the extension of the CVC-ColonDB database. In the original experimentation, non-overlapping images are kept in training and testing sets. The CVC-ColonDB was split, 80% was kept for training, and the rest was used to validate the proposed model. We achieved a DSC of 0.867 and an F1 score of 0.789 on the validation set consisting of 60 images. Similarly, On the ETIS-Larib database, we achieved a DSC and F1 score of 0.801 and 0.786, respectively, on the validation sets with a similar split. Different shapes of the polyps will have a negligible effect on the segmentation performances by our proposed approach. Our proposed model uses both the saliency map and shape compactness for the polyp segmentation. Most of the polyps have circular or ellipse shapes. Therefore, the proposed SC prior to the saliency map works well. Figure 3.21 and Figure 3.22 demonstrate the proposed approach's segmentation performances on some of the colonoscopy frames. The DSC for



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

Figure 3.23: Polyp segmentation on some example samples with different polyp shapes; (a), (c), and (e) Raw images, (b) (d) and (f) corresponding generated segmentation mask.

Figure 3.23b, 3.23d and 3.23f are 0.93, 0.91, and 0.871 respectively.

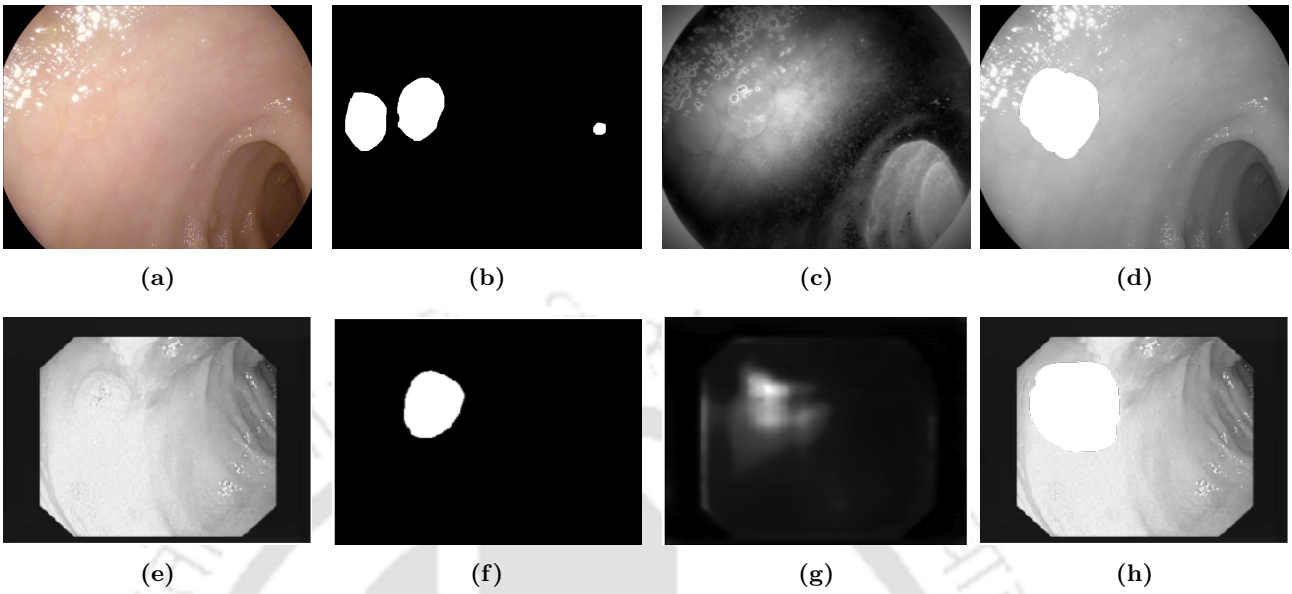


Figure 3.24: Improper polyp segmentation in some example samples of both the datasets; (a)(e) raw images, (b)(f) ground-truth masks, (c)(g) saliency maps, (d)(h) obtained segmentation results.

To provide a fair comparison of our method, we compare our results with some of the most recent state-of-the-art methods. Most of the methods in the literature used Narrowband image (NBI) images. Texture and other features are more visible in NBI polyp images [243,244]. Methods used in [161,177] use NBI images. Other methods in the table used the CVC-ClinicDB database for validation of their methods. A direct comparison of our results with the state-of-the-art methods is given in Table 3.7 and 3.8. Some qualitative performances of the proposed method is given in Figure 3.22. The proposed method achieves a mean Dice value of 82.11%, which is very competitive with the state-of-the-art techniques. Our proposed model fails when the polyp is visually indistinctive from the background. Figure 3.24 shows the poor segmentation results obtained by our proposed method. The polyps are texturally indistinctive from the background. The role of the shape compactness prior is negligible in these two cases. Therefore, our proposed approach works well when both the texture and shape are visually distinctive.

Statistical analysis is carried out to validate the effectiveness in performance of our approach compared to the base models. This involves comparing the histogram of the polyps' ground-truth (GT) masks with the obtained masks by our proposed method. In our analysis, Chi-square (χ) distance is adopted to quantify the matching between two histograms. Let h_1 and h_2 be two histograms, the χ

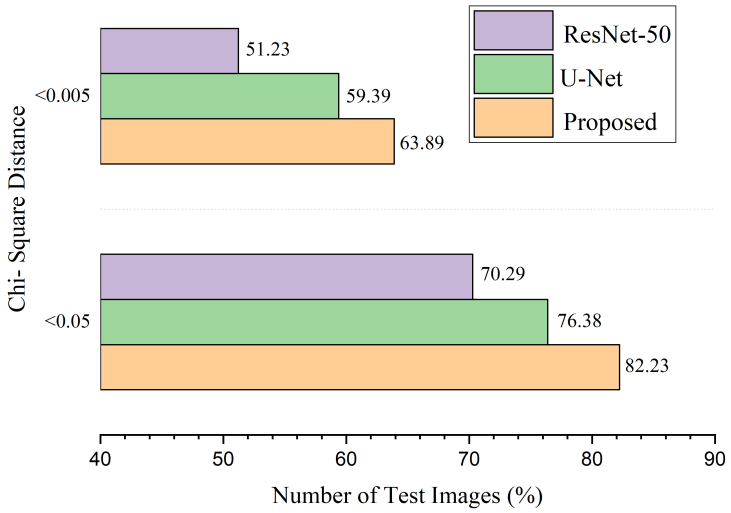


Figure 3.25: Comparative distribution of number of images vs Chi-square values obtained using histogram comparison.

is computed as:

$$\chi(h_1, h_2) = \sum_I \frac{(h_1(I) - h_2(I))^2}{h_1(I)} \tag{3.36}$$

The low value of the metric corresponds to a better match of the two histograms. The χ values are divided into two ranges. Figure 3.25 shows that 63.89% of the test images significantly match with their corresponding ground truths and attain a χ distance of less than 0.005. However, only 59.39% of the test images attain such χ range for the U-Net. Similarly, in the case of the Resnet50 model, only 51.23% of the test images achieve such χ range. From Figure 3.25, it can be concluded that our method is better suited for polyp segmentation compared to the base models.

3.4.3 Conclusion

In our method, shape compactness prior is used to detect the ROI of arbitrary polyp shapes. Salient object detection is a method to accentuate the object of interest. A deep U-Net-based CNN model is used for saliency detection in our approach. Segmentation is formulated by energy terms given by a probability map and compactness prior. To solve the energy function, ADMM is used, which is efficient and fast. A series of comparative tests using the publicly available datasets of colonoscopy polyps are used to assess the effectiveness of our method. Compared to the state-of-the-art methods, the experimental results suggest that the proposed technique has competitive performance. Our

segmentation method can be used for colonoscopy image analysis. In the future, other priors can also be considered for our model to detect salient regions and may help in better polyp detection.

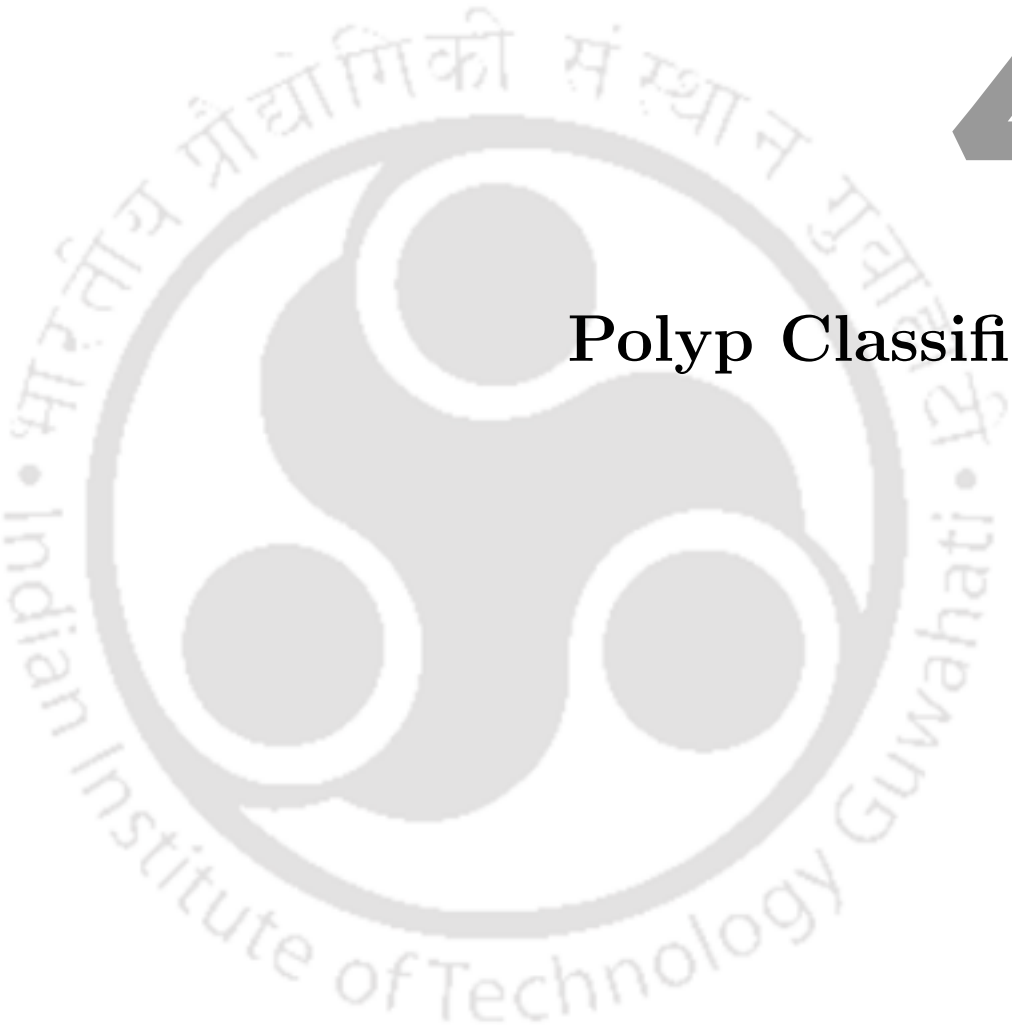
3.5 Summary

Three methods are proposed in this chapter to solve the segmentation problem for different scenarios. The first approach can select clinically relevant frames followed by the segmentation of polyps. This technique provides a preliminary study on the MDE. This method applies to the advanced staged large polyps needing immediate clinical attention. However, the small and patchy polyps may progress to malignancy over time. The analysis of these polyps is also crucial for early diagnosis. Therefore, an MRF-based technique is proposed, utilizing the polyp pixels' color, texture, and spatial information to segment different polyp structures. However, textureless and nascent polyps may sometimes get unsegmented with this method. The polyp shape information is also incorporated with the extracted deep features to leverage this.

The clinicians comprehensively analyze the detected polyps to find cancer in them. However, the similarity in pathological manifestations across diseases makes polyps' manual inspection and annotation cumbersome and inefficient. The polyp characteristics may not always be visible to the human eye, and diagnostic information may be ignored, making the wrong diagnosis. To salvage this, An automated polyp classifier for polyp classification, i.e., adenoma (malignant) and hyperplastic (benign), is provided to solve the problems mentioned above. Further, a method is devised for the automated grading of dysplasia on histopathological polyp images. The details of the methods are discussed in Chapter 4.

4

Polyp Classification



Contents

Objective

During the colonoscopy, the doctors extract the polyp regions. Then, they analyze different characteristics of these regions, such as geometry (shape and size), the surface of the polyp (texture), color (blood traces), boundary (smooth or wavy), etc., to classify them into different grades of carcinoma. Sometimes such features, especially the texture features are not visible in the frames, making it very difficult to decision making. Sometimes the diagnostic information gets overlooked because of susceptibility to the error of omission. Manual inspection of a large number of frames needs enormous effort and time. In this chapter, automated polyp classifiers are proposed for two-class polyp classification, i.e., adenoma (cancer) and non-adenoma (non-cancer), to circumvent the above difficulties. In the first approach, the shape and texture features are effectively extracted to discriminate the polyps types. To extract features from the small and imbalanced dataset effectively, we adopted a similarity learning approach based on the Triplet network in our second approach. Further, fusing the shape features with the extracted embeddings from the Triplet network enhances the classification performance. The last work in this view is based on the classification of polyps from the polyp histopathological images, which are generally used in a clinical setup.

4.1 Introduction

Texture feature analysis is very important in polyp classification. The texture descriptors are capable of characterizing the polyps. In this work, we are concerned about polyp classification, and hence literature study is confined only to this domain. Handcrafted feature learning-based methods for classification are inevitable when the dataset is in paucity. In [124], Stehle et al. used vascularization features for colon polyp classification in endoscopic images. Condessa et al. [125] used curvature features for detection and classification of colorectal polyps. Fu et al. [126] used texture features taken from both the spatial domain and spectral domain. They applied principal component transform (PCT) and extracted the texture features from the first component of the PCT. Reduced feature dimension set was prepared using sequential forward selection (SFS) and sequential floating forward selection (SFFS) algorithms for classification using SVM. Hafner et al. [127] proposed local texture properties using 1D histogram using similarity of neighboring pixels. The compact colour vector features were classified using the k-nearest neighbor classifier. Mesejo et al. [128] used combination of features for better representation of polyp features. They used texture features and 3D features for

classification of polyps. In [129], Wimmer et al. used wavelet-based features for polyp classification. Engelhardt et al. [130] proposed Color-GLCM features and SVM classifier for the task. In [131], Bag of words (BOW) descriptors with spatial pyramid matching (SPM) were used. Some deep-learning-based methods along with their performances are elaborately discussed in section 4.2.2. These methods discussed above mainly used global features for polyp classification. In most of these methods, a particular feature descriptor is used which may not be sufficient for the complete characterization of polyp features. Also, Most of the methods are experimented on a very limited dataset. The performances of these methods are not satisfactory. The rest of the chapter is organized as follows. A local shape and texture features based approach is discussed in section 4.2. Section 4.3 describes a feature fusion based approach for polyp classification. Section 4.4 describes our third standalone approach on the proposed GAN model. Finally, the summary of the chapter is provided in section 4.5.

4.2 Local Shape and Texture Features for Classification

In our earlier work [174], we used Gabor filter banks in the initial stage, followed by discrete cosine transform (DCT) upon the sub-bands. The square of the DCT coefficients calculated on all the Gabor filter sub-bands is considered the energy. The DCT coefficients derived from all the sub-bands are fused to form the feature vector. Local binary pattern (LBP) as texture features is also used. The malignant polyps are expected to have more texture on their surface than the benign polyps. However, the method fails to characterize the polyp features efficiently. An accuracy of 74.03% was achieved with this method. The texture features are more prominent on different scales; therefore, multiscale feature representation can better characterize the polyp texture features. Polyp shapes impart discriminant features among the polyp classes. Therefore, this feature can also be utilized along with the texture features.

In our proposed method[1], the texture feature imparted by the surface roughness is described by the fractal weighted local binary pattern (FWLBP) descriptor. The FWLBP is a variant of LBP, where the histogram is given by LBP, weighted by the fractal dimension (FD). The shape and irregularity of the surface are quantified by the pyramid histogram of oriented gradient (PHOG) features. The PHOG features represent the local shape by using a histogram of oriented gradient (HOG). For the selection of important features, a fuzzy entropy-based algorithm is deployed. Support vector machines (SVM)

[1]This work has been published in IEEE Access, 2021 (Refer *List of publications* page for details).

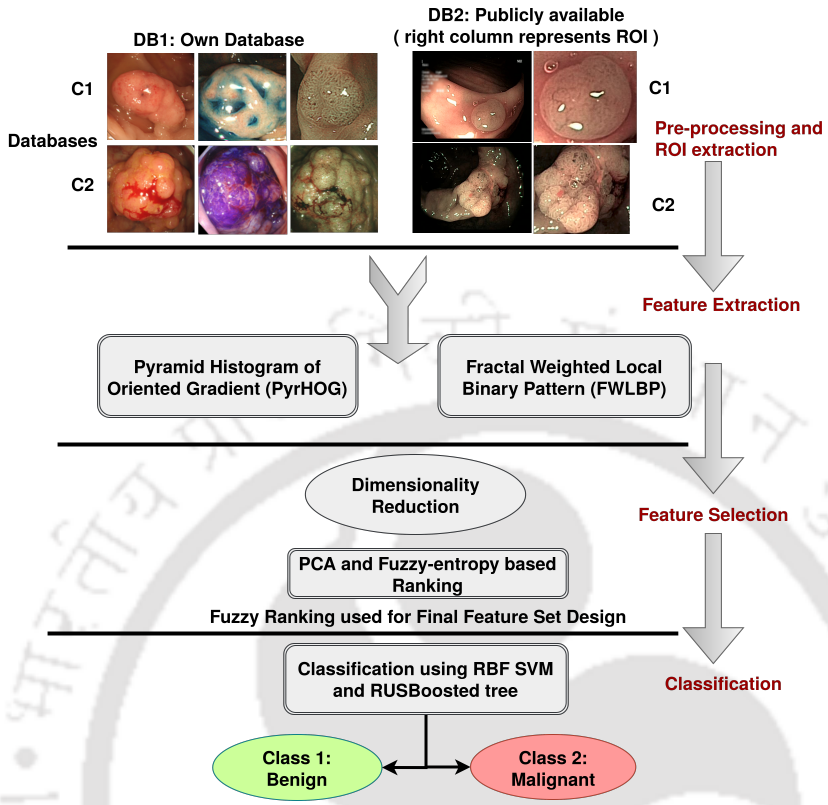


Figure 4.1: The proposed framework (DB1: database 1 and DB2: database 2, C1: Class 1 (benign) and C2: Class 2 (malignant)). DB1 contains NBI, Dye and white light images, whereas DB2 contains NBI and whitelight images.

and RUSBoosted tree classifiers are used for polyp classification. K-fold cross-validation technique is adopted to avoid over-fitting. The proposed method can do real-time polyp classification. The feature descriptors employed in our work are robust to affine transformation and mild illumination variations. Our proposed method gives better performance on samples taken from all the endoscopic modalities. Endoscopic videos are captured under different geometrical positions/transformations and lighting conditions, and our proposed feature representation scheme shows the robustness and adaptation for all these cases. The rest of the work is organized as follows – Section 4.2.1 gives the schema of the proposed work. Results and discussions are explained in section 4.2.2. Finally, section 4.2.3 concludes the work.

4.2.1 Proposed Method

All the steps and methods applied in our proposed method are briefly discussed in the following sub-sections. The proposed framework is shown in Figure 4.1.

4.2.1.1 Pre-processing and ROI extraction

Classification of polyps requires some pre-processing, in which region of interest (ROI) is extracted from each of the frames. The frames from the endoscopic video sequence are detected, which have at least one polyp. Then, the ROI can be segmented automatically or manually. In this work, the polyp regions are segmented manually by an expert in this domain.

4.2.1.2 Shape descriptor PHOG

The shape is an important attribute in detecting malignancy in polyps. The shape of the polyp boundary, irregularities in the surface of the polyp region, etc., are general shape attributes. The shape features are generally classified into two categories– region-based and contour-based. The region-based shape descriptor encapsulates the shape information by considering all pixels in a region. On the contrary, the contour-based shape descriptor only provides shape information of the boundary of the ROI.

In our method, a region-based descriptor is used. The local shape of the polyp is captured based on distribution over edge orientations within a region and spatial layout. This information is obtained by tiling the image into regions at multiple resolutions based on spatial pyramid matching [245]. The descriptor consists of a histogram of orientation gradients over each image subregion at each resolution level - a pyramid of histograms of orientation gradients (PHOG). The histogram of edge orientations within an image subregion encapsulates the local shape information. The contribution of each edge is weighted according to its magnitude, in a manner similar to SIFT [246]. Each bin in the histogram represents the number of edges that have orientations within a certain angular range. This representation for regions is referred to as histogram of orientated gradient (HOG) [247]. Spatial pyramid matching [245] (SPM) is used to incorporate spatial information in the HOG vectors. The HOG vector is computed at each pyramid resolution level (L). Finally, all the HOG vectors are concatenated to represent the PHOG descriptor. The dimensionality of the PHOG descriptor for the entire image is given as:

$$K * \sum_{l=0}^{L} 4^l \tag{4.1}$$

The histograms of the same level are concatenated into one vector. All vectors at each pyramid resolution are concatenated to get the final PHOG descriptor. In our case, we set K = 8 and L = 3. Hence, our PHOG descriptor is of length $8 \times (1 + 4 + 16 + 64) = 680$. The bins are angular

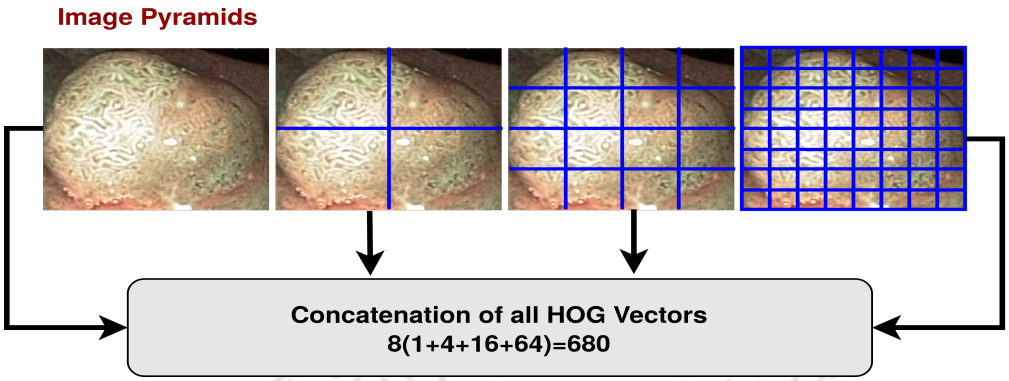


Figure 4.2: Extraction of PHOG features from a colonoscopic polyp image sample. Grids at three pyramid resolution in the original image; concatenation of all the HOG vectors in three pyramid resolutions to obtain the PHOG features of a sub-image.

orientation of the edge features calculated using canny edge detector in our method. The HOG vector is calculated at each pyramid resolution level (L). Figure 4.2 describes the HOG vector calculated at each pyramid level. The histogram of edge orientations within an image sub-region encapsulates local shape information. The PHOG descriptor is normalized to sum to unity. This ensures that texturally rich images with more edge strength, or are larger, are not weighted more strongly than others.

4.2.1.3 Texture and shape decriptor FWLBP

Texture patterns of a polyp play an important role in classification. The doctor analyzes the surface irregularity and roughness of a polyp before taking any pathological interpretation. An ideal model must be robust to view-point alteration, illumination variations, rotation, reflection, scale change, and geometry of the underlying surface. All the mentioned cases are commonly encountered during colonoscopy. In [248, 249], local texture descriptors are used to achieve local invariance. The multi-resolution analysis represents image features as coefficients along multiple scales, directions, and resolutions [250–252].

The FWLBP descriptor is robust to scale, rotation, and illumination variations. The invariance property of FWLBP is due to an interesting characteristic shown by the fractal transformation, which is invariant under *bi-lipschitz transform*. Any traditional transform (like translation, rotation, scaling, and viewpoint change) is a *bi-lipschitz transform* [253]. During a colonoscopic procedure these transformations are commonly encountered in different polyp frames, and the descriptor should be robust to all these transformations. Our proposed FWLBP is robust to these transformations. To experimentally validate this, we manually introduced all the transformations in the polyp frames, and

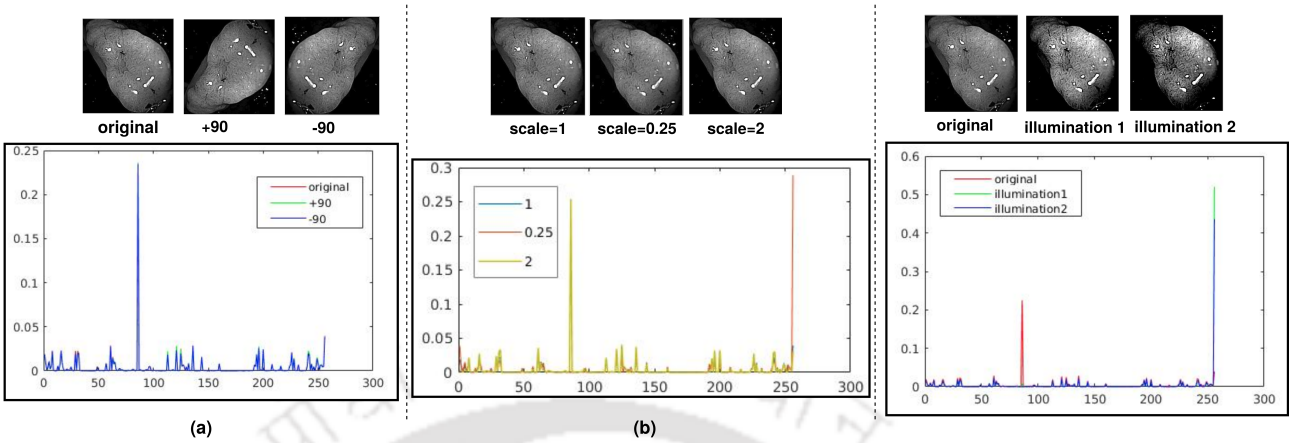


Figure 4.3: Illustration of the robustness of FWLBP histograms to different image transformations; (a) Rotation (b) Scale (c) Illumination.

subsequently, the FWLBP features are extracted. The histogram plots of the original frames and the respective transformed images are analyzed, and it is seen that the histogram plots of the transformed images are identical to the respective histograms of the original polyp frames. Song et al. [254] used histogram plots to illustrate the invariance properties of their proposed feature extractor to the transformations. Similarly, Roy et al. [255] employed histogram plots to elucidate the invariance property of their proposed features. Figure 4.3 shows different transformations of polyps and the corresponding histograms of the extracted FWLBP descriptors. This analysis shows that FWLBP is robust to different image transformations.

A texture can be defined as a geometric multiscale self-similar structure like a fractal dimension (FD) or LBP. In our proposed work, the FWLBP descriptor is used as the feature descriptor, as it can give a promising performance under scale, rotation, reflection, and illumination variation. Also, the fractal model can be used to obtain shape information [256]. The fractal dimension in the image possesses the characteristics of the self-similar texture pattern. It characterizes the roughness and shape of the geometry. Let us consider a bounded set B in an n-dimensional euclidean space S. The set B is said to be self-similar when B is the union of N_r, distinct (nonoverlapping) copies of itself each of which is similar to B scaled down by a ratio r. It can be written as [257]:

$$1 = N_r(b)r^{D(b)} \text{ or } D(b) = \frac{logN_r(b)}{log\frac{1}{r}} \tag{4.2}$$

where, $D(b)$ is a density function. The density function represents the fractal dimension (FD). For a 2-D structure, if we magnify it by $R = 2$, we would get $N_r = 4$ ($N = R^2$) copies of the 2-D structures.

Similarly, if we take R=3, we would get 9 structures *i.e.,* $N_r = R^2$. Thus, without loosing generality, for a D dimensional figure, if the magnification factor is R than the resultant copies will be $N = R^D$. Taking log on both sides, we would get D=log$N$/log$R$, which is represented in Eq. (4.2), where R =$1/r$ of Eq. (4.2) and N_r is the number of identical or self duplicating copies of the fractal. From Eq. (4.2), it is difficult to calculate D directly. Several methods have been proposed, but our work adopts a differential box-counting (DBC) method to find the FD images. Before applying DBC, the images are stacked into pyramids using Gaussian scale-space with variance r to achieve scale-invariance as per Eq. (4.3).

$$\Lambda_{x,y,r} = G_r(x, y, r) * I(x, y) \tag{4.3}$$

where,

$$G_r(x, y, r) = G_r(x)G_r(y)$$

$$G_r(x) = \frac{1}{r\sqrt{2\pi}}\exp\frac{-||x||^2}{2r^2} \text{ , and } G_r(y) = \frac{1}{r\sqrt{2\pi}}\exp\frac{-||y||^2}{2r^2}$$

After constructing the Gaussian scale space, the DBC algorithm is applied to each layer of the scale space. The resultant intermediate images are combined into the final FD images using a linear regression technique [258]. Let the image be represented by I(X×Y). The Gaussian pyramid $\Lambda_{x,y,r}(X \times Y \times L)$ is generated from I(X×Y) with L levels with scaling factor ranging from $[r_{\min}.....r_{\max}]$. The DBC algorithm is implemented by applying a variable size non-linear kernel to image cells with $f_{\max}$ and $f_{\min}$ are maximum and minimum intensity values. The kernel $k_{i,j}(m \times n)$ is given by Eq. (4.4):

$$k(i, j) = \sum_{i=-\alpha}^{\alpha}\sum_{j=-\beta}^{\beta}\left\lfloor\frac{f_{\max} - f_{\min}}{r}\right\rfloor + 1 \tag{4.4}$$

where, r is the scaling factor of the layer l, and

$$\alpha = \left\lceil(\frac{m-1}{2})\right\rceil, \text{ and } \beta = \left\lceil(\frac{n-1}{2})\right\rceil \tag{4.5}$$

$\lfloor\cdot\rfloor$ and $\lceil\cdot\rceil$ are floor and ceil functions. α and β are used to centre the kernel on the pixel $f_{x,y}$. The output of the operation is given as:

$$\Gamma(x, y, r) = \sum_{i=-\alpha}^{\alpha}\sum_{j=-\beta}^{\beta} k(i, j)\Lambda(x + \alpha, y + \beta, r)\left(\frac{L}{r}\right)^2 \tag{4.6}$$

where, $\Gamma_{(x,y,r)}$ represents a matrix of intermediate images consisting of layers formed with kernel

varying from r_{min} to r_{max} and is given by:

$$\Gamma(x,y,r) = \begin{bmatrix} g_{11r} & g_{12r} & \cdots & g_{1Yr} \\ g_{21r} & g_{22r} & \cdots & g_{2Yr} \\ \vdots & \vdots & \ddots & \vdots \\ g_{X1r} & g_{X2r} & \cdots & g_{XYr} \end{bmatrix}$$

The pixel value $F(x,y)$ of the FD image is given by the slope $m_{x,y}$ of the linear regression line between $\Gamma_{(x,y,r)}$ and r. Let Φ be a vector, where element θ_1 corresponds to first gray values of all layers and so on.

$$\Phi = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_{X\times Y} \end{bmatrix} = \begin{bmatrix} g_{111} & g_{112} & \cdots & g_{11L} \\ g_{121} & g_{122} & \cdots & g_{12L} \\ \vdots & \vdots & \ddots & \vdots \\ g_{XY1} & g_{XY2} & \cdots & g_{XYL} \end{bmatrix}$$

Then, the slope is calculated as follows:

$$\lambda_1 = \sum r^2 - \frac{\left(\sum r\right)^2}{L} \tag{4.7}$$

$$\lambda_2 = \sum r\Phi - \frac{\left(\sum r\right)^2\left(\sum \Phi\right)^2}{L} \tag{4.8}$$

Finally, the FD image $F_{x,y}(X \times Y)$ is given by:

$$F(x,y) = m_{xy} = \sum_{x=1}^{x=X}\sum_{y=1}^{y=Y}\frac{\lambda_2}{\lambda_1} \tag{4.9}$$

Feature representation using FD and LBP: Since the fractal dimension is a logarithmic function (Eq. (4.2)), it deals with the effects of illumination, as the log function expands the values of darker pixels and compresses the brighter pixels in the image. Thus, the proposed descriptor is insensitive to scale, translation, rotation or reflection, and illumination change.

The entire process of feature set generation is shown in Figure 4.4. The FD images are generated using different scales with the DBC technique [259] (Figure 4.4(b)). Figure 4.4(c) shows the generation of LBP images using three sampling radius (1, 2, and 3) with $N = 8$, the number of samples. The LBP value at pixel location (x,y) for a given R and N is given by:

$$\mathrm{LBP}_{x,y}^{N,R} = \sum_{n=1}^{N} 2^{n-1} \times \mathrm{sign}\left(J_{x,y}^{(N=8,R)} - \alpha(x,y)\right) \tag{4.10}$$

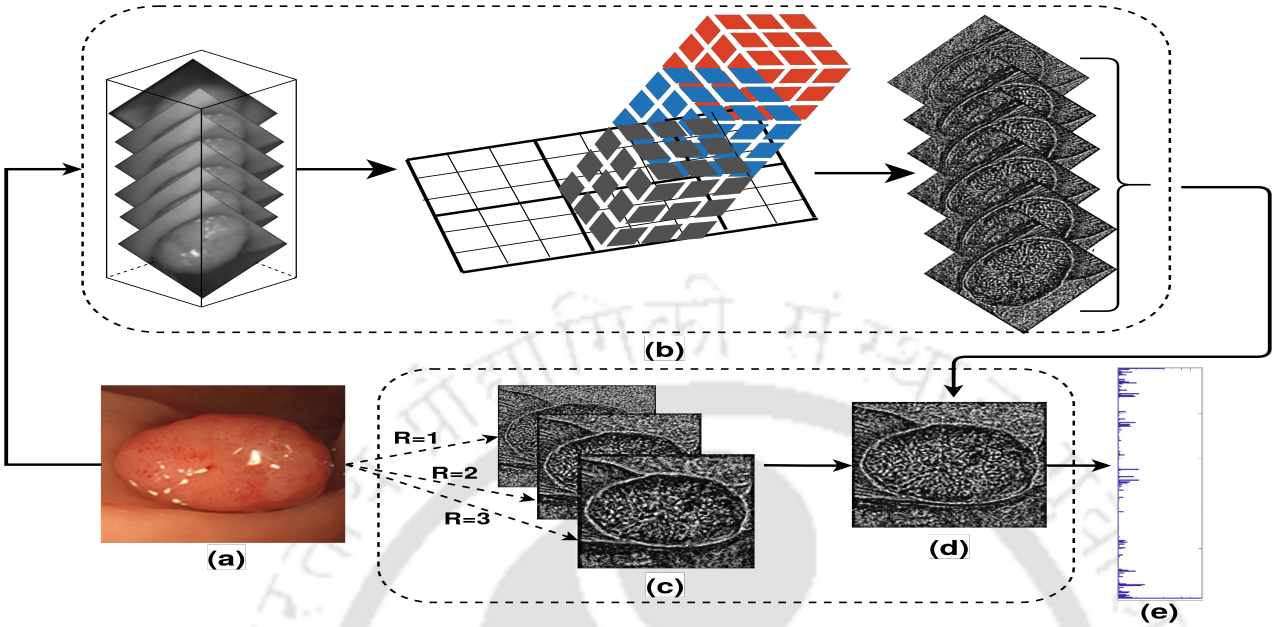


Figure 4.4: Texture feature extraction using the proposed FWLBP (a) Input colonoscopic frame (b) Fractal dimension (FD) image using spatial pyramid and DBC algorithm (different boxes are represented by different colours) (c) LBP calculation for different sampling radius (R) (d) Weighted LBP image formed using FD weights (e) Final feature representation.

where, the centre value $\alpha(x, y)$ of a mask of radius R is placed at location (x, y) with neighbour pixel values given by $J_{x,y}$. To make the calculation simple, the value of N is taken as 8 with uniform sampling distance.

This forms the LBP images. An indexing operation is performed to combine already formed FD images with the LBP histogram to create the final histogram. The algorithm looks for the locations in the LBP images $\text{LBP}_{x,y}^{N,R}$ which have the same pixel values $j_{x,y}^{N,R}$. Then, it finds the FD values in the corresponding locations of FD images and sums them up. The sum is the weight of the value $j_{x,y}^{N,R}$ in the histogram.

$$\text{FWLBP}^{R,N}(j) = T\left(\text{LBP}_{x,y}^{N,R}, F_{x,y}, p\right), j \in [0, 2^{N-1}] \tag{4.11}$$

where,

$$T\left(\text{LBP}_{x,y}^{N,R}, F_{x,y}, j\right) = \begin{cases} \sum F_{x,y}, & \forall\ LBP_{x,y}^{N,R} = j \\ 0, & \text{otherwise} \end{cases}$$

where, $\text{FWLBP}^{R,N}(j)$ gives weight j in the histogram. $T(\cdot)$ is an indexing operation to combine FD and LBP. For final feature construction, all $\text{FWLBP}^{R,N}(j)$ are concatenated. The length of the feature vector is of 768 dimensions $(256 + 256 + 256)$ with $R = 1, 2,$ and 3 with six levels of image pyramids

Table 4.1: Initial parameter settings for PHOG and FWLBP

Descriptor	**Bin size (K)**	**Pyramid level (L)**	**Final feature dimension (F1)**
PHOG	8	3	680
Descriptor	**Radius size (R)**	**Pyramid level**	**Final feature dimension (F2)**
FWLBP	1, 2, 3	[2-7]	768

$[r_{min} = 2\ ,\ r_{max} = 7]$

In our proposed work, two feature representation schemes are employed. One using the PHOG feature with dimensions 680, the other descriptor is FWLBP with feature dimensions 768. The final feature is given by the concatenation of the feature sets. The features and the parameters of the descriptors are given in Table 4.1.

4.2.1.4 Feature selection

Selection of the most discriminative features is very important in classification. In this work, the final feature set is selected using a possibilistic method. A feature ranking algorithm based on fuzzy-entropy for the selection of the final feature set is adopted in our work. In this approach, the features are grouped into different sets at first, and subsequently their contribution to an assigned class is evaluated [260]. The groups are then ranked according to their performances. Thus, it helps in the selection of optimal features for classification. The selection is based on mutual information (MI) between a particular feature (f) and class label (C). Fuzzy based feature ranking is performed in this work. In doing so, fuzzy entropy based feature selection is applied for ranking the features. Based on the ranking, feature dimensions of size k = 34, 68, 102, . . . , m have been designed, where m = 680 for PHOG and k = 48, 96, 144, . . . , 768 for FWLBP. The minimum number of features are judiciously selected on the basis of large number of experimentations, and numbers are 34 and 48 for PHOG and FWLBP, respectively. Feature dimension below these numbers would yield low accuracies. Unlike other feature selection techniques like PCA, ICA, etc., our proposed method ranks the features based on their contributions to an assigned class. The details of the technique is given below:

Let us consider a feature vector $FV = \{f_1, f_2, ..., f_l\}$, where l is the number of features. Fuzzy membership value that k^{th} vactor will be in i^{th} class is calculated as follows:

$$\mu_{ik} = \left(\frac{\|f_i - f_k\|_\sigma}{r + \epsilon} \right)^{\frac{-2}{m-1}} \tag{4.12}$$

where, m is the fuzzification parameter and ϵ is a value to avoid singularity and σ is the standard deviation. Finally, the membership of each of the samples in all the classes is normalized according to

$$\sum_{i=1}^{c} \mu_{ik} = 1.$$

In case of total c numbers of classes, c numbers of fuzzy sets along each feature f needs to be considered. Each of these reflects the membership degree in c problem classes. Fuzzy joint probability of a particular feature vector belonging to a class c can be given by the formula:

$$P(f, c_i) = \frac{\sum_{k \in V_i} \mu_{ik}}{N} \tag{4.13}$$

where, $P(f, c_i)$ gives the degree of contribution of a feature to a particular class. V_i indicates the indices of the feature vector that belong to class i and N indicates the total number of patterns or the dimension of feature vector. The joint fuzzy entropy of features of each class can be calculated as follows:

$$H(f, c_i) = -P_{f,c} log P_{f,c_i} \tag{4.14}$$

The complete fuzzy entropy can be obtained as:

$$H(f, C) = \sum_{i=1}^{c} H(f, c_i) \tag{4.15}$$

Marginal entropy $H(f)$ can be found as $H(f) = -P_{f_{X_i}} log P_{f_{X_i}}$, where X_i indicates the c numbers of fuzzy sets and $P(f_{X_i})$ can be calculated as:

$$P(f_{X_i}) = \frac{\sum_k \mu_{ik}}{N} \tag{4.16}$$

Marginal class entropy $H(C) = -P_{c_i} log P_{c_i}$. Then, mutual information (MI) between particular feature and class label can be calculated using the formula:

$$MI(f; C) = H(f) + H(C) - H(f, C) \tag{4.17}$$

4.2.1.5 Classification

The feature dimensions of the proposed descriptors are very high and the features are not linearly separable in the feature space. Kernel-based Support vector machine (SVM) [261] can perform well in such scenarios. For the SVM classifier, 10-fold cross-validation techniques are adopted for training and testing. The MLP is trained with 70% of total samples and 15% samples are kept each for testing and validation. The open database (DB2) contains a very small number of samples with a large class imbalance. RUSBoosted tree is used as the classifier for the database, as it can handle the class imbalance [262]. The details of the databases are given in section 4.2.2.

4.2.2 Results and discussion

The proposed work is validated with two databases, one is an annotated dataset by a medical practitioner, and the other is publicly available. Our generated dataset (DB1) was developed by the Department of Gastroenterology, Aichi Medical University Hospital, Nagakute, Japan, under the supervision of Dr. Kunio Kasugai. The images were acquired with the consent of the subjects and by following proper ethical protocols. Strict guidelines and regulations were followed while generating the database. He classified the acquired polyps into two classes, namely malignant and benign. The Aichi Medical University ethical committee has approved this clinical study (January 15, 2018; Approval No. 2017-H304). The institutional review board of Aichi Medical University Hospital approved the collection of the data. The other dataset (DB2), a colonoscopy video dataset [128] is publicly available in

Table 4.2: Details of the datasets. C1: benign and C2: malignant

Database	Type	Class	Total frames
DB1	NBI, WL, and Dye	C1	373
		C2	208
DB2	NBI and WL	C1	NBI-21/WL-21
		C2	NBI-40/WL-40

the url: `http://www.depeca.uah.es/colonoscopy_dataset/`. DB2 consists of video sequences from three classes, namely, adenoma (malignant), hyperplasia (benign), and serrated (intermediate stage). We only considered the adenomatous and hyperplastic polyps in order to compare with DB1. The details of the datasets are given in Figure 4.5 and Table 4.2. The ROIs were manually segmented from the frames by the experts.

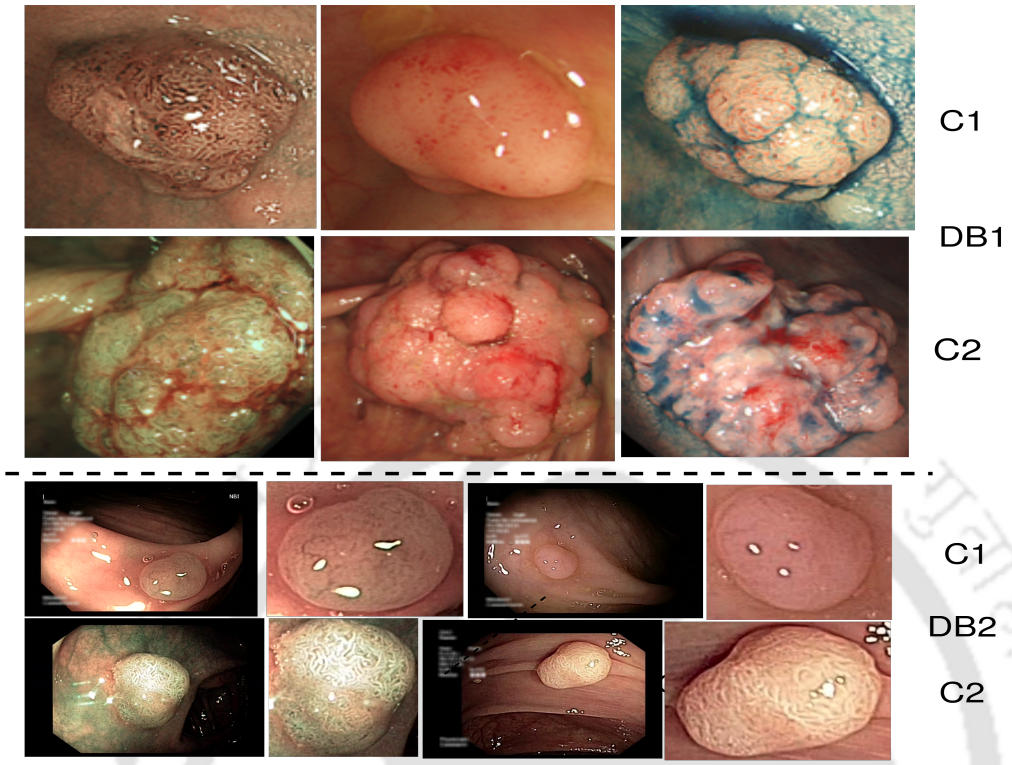
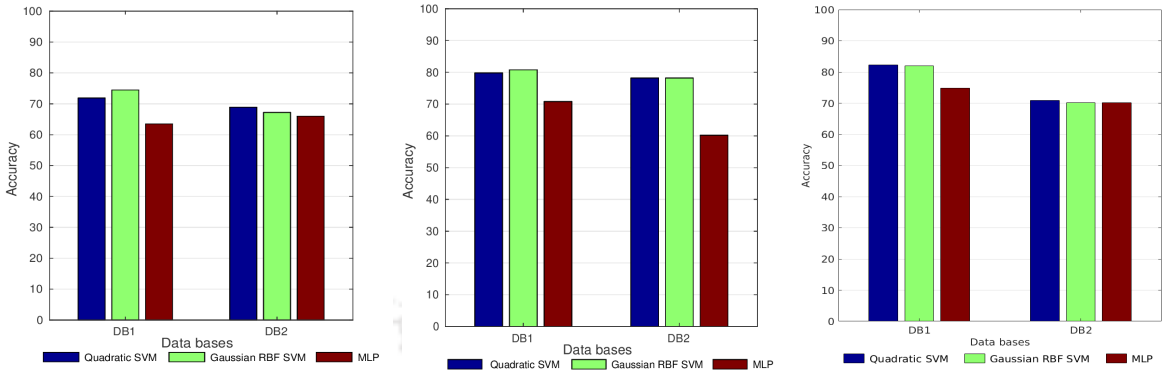


Figure 4.5: Dataset with sample images (C1: benign and C2: malignant samples): Sample frames of both the classes; In DB1: 1^{st}, 2^{nd}, and 3^{rd} column samples are NBI, WL and Dye images, respectively. In DB2: 1^{st} and 3^{rd} images of each row are NBI and WL, respectively, and 2^{nd} and 4^{th} images are the ROI of the corresponding frames.

4.2.2.1 Feature set design and final classification

Let $\boldsymbol{F_1}=\{\boldsymbol{f_1}, \boldsymbol{f_2}, ..., \boldsymbol{f_{680}}\}$, and $\boldsymbol{F_2}=\{\boldsymbol{f_1}, \boldsymbol{f_2}, ..., \boldsymbol{f_{768}}\}$ represent feature vectors for PHOG and FWLBP, respectively. Initially, classifications are performed for the two databases considering all the feature dimensions of the two descriptors. Individual performance and performance after combining $\boldsymbol{F_1}$ and $\boldsymbol{F_2}$ are also analyzed. Results are shown separately for shape features, PHOG $\boldsymbol{F_1}$ and texture and shape features, FWLBP $\boldsymbol{F_2}$ for both the datasets. Polynomial kernel-based SVM (Quadratic SVM), Gaussian RBF SVM, and MLP classifiers are used in the assessment of classification performances. An analysis of classifier performances are given in Figure 4.6 and Figure 4.9.

MLP was trained with 70% samples and 15% samples were kept each for testing and validation. For DB2, 50% of total samples were used for training and 25% each for testing and validation. Using PHOG feature, the average best accuracy is given by RBF SVM, which is 74.5% and 67.2% for DB1 and DB2, respectively. Similarly, the result for FWLBP is 80.8% and 78.2% for DB1 and DB2, respectively. With the combination of both features, the performance is improved for both datasets.



(a) (b) (c)

Figure 4.6: Accuracies using proposed features for different classsifiers (a) PHOG (b) FWLBP (c) PHOG+FWLBP.

Table 4.3: Final classification results for DB1

	Original feature dimension	Reduced feature dimension	Accuracy
PHOG	680	170	78.6
FWLBP	768	432	82.2
PHOG+FWLBP	1448	800	84.1

Then, we introduced some feature selection methods. Also, from the above analysis, it is clear that SVM performance is better than MLP across the databases.

The statistical significance test is generally performed in big data analysis to know the contribution of attributes along different independent factors. Thus, the significance of each of the variables in the two feature vectors $\boldsymbol{F_1}$ and $\boldsymbol{F_2}$ along the two classes are studied. ANOVA (Analysis of variance) test was conducted where the values of the features are taken as the dependent variables (DVs) and class as a constant factor. The last column i.e., *sig* of the output ANOVA test result is analyzed. If this value is less than equal to 0.5, then the particular variable seems to be contributing more in the

Table 4.4: Final classification results for DB2

	Original feature dimension	Reduced feature dimension	Accuracy (RBF SVM)	Accuracy (RUSBoosted Tree)
PHOG	680	136	85.2	87.2
FWLBP	768	384	80.2	81.8
PHOG+FWLBP	1448	500	86.6	90.16

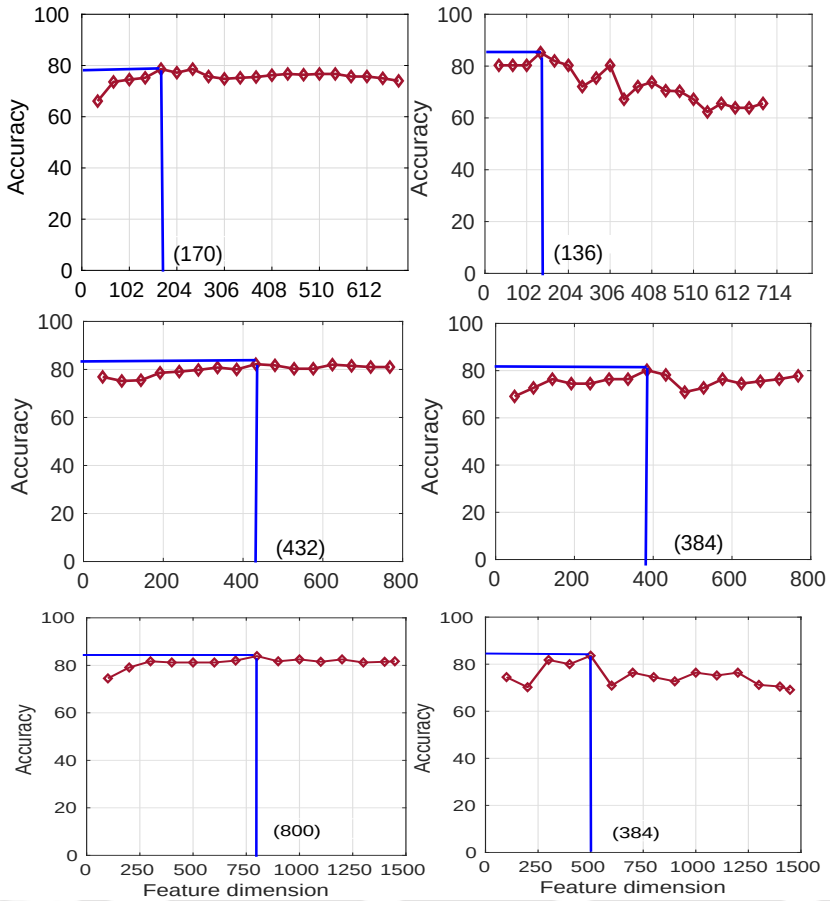


Figure 4.7: Fuzzy entropy based feature selection; first row left PHOG on DB1, and right PHOG on DB2, middle row left FWLBP on DB1, and right FWLBP on DB2, bottom row left PHOG+FWLBP on DB1, and right PHOG+FWLBP on DB2.

classification according to the null hypothesis. The statistical significance test using one-way ANOVA was done using GNU PSPP, version 0.8.5-5. The result shows that with FWLBP for DB1, most of the variables reject the null hypothesis with *sig.* value, p=0.000. The PHOG features also have a similar kind of response. Similarly, ANOVA for both the feature representation schemes $\boldsymbol{F_1}$ and $\boldsymbol{F_2}$ showed that the overall influence of DVs are highly significant for PHOG and FWLBP for DB2. For DB2, ANOVA for PHOG descriptor gives a better result than FWLBP, where, p=0.016 for the FWLBP and most of the variables do not satisfy the null hypothesis. The test suggests that both the feature representation schemes are important and it suggests the requirement of feature optimization, i.e., feature selection. The feature selection based on fuzzy entropy is given in Figure 4.7. The change in feature dimension and accuracy are given in Table 4.3 and Table 4.4. The performance using AUC for both the datasets is given in Figure 4.8

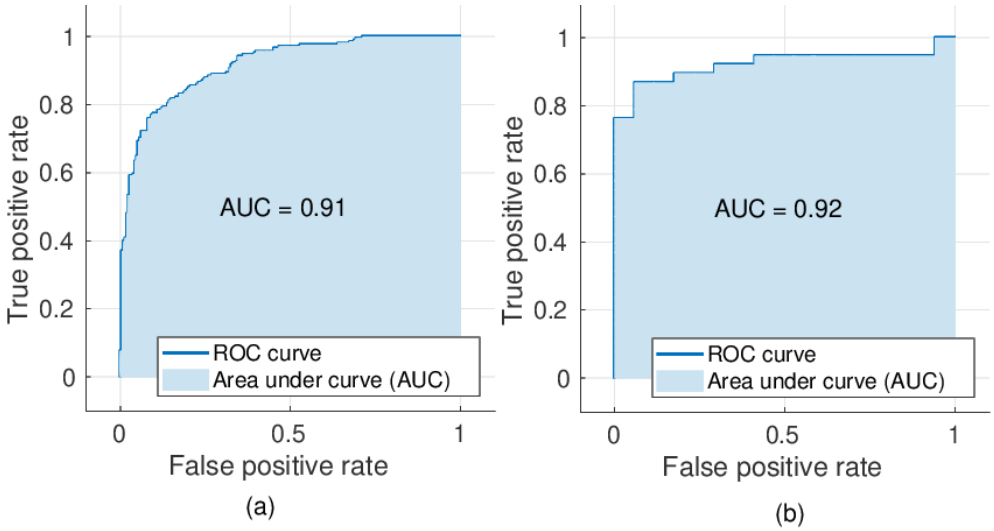


Figure 4.8: Performance analysis using AUC for both the databases (a) DB1 (b) DB2.

4.2.2.2 Comparision

Deep learning-based methods have been recently used in medical image and video processing. Many researchers have recently started deploying deep models for polyp classification. These techniques generally require a large number of labeled training samples. However, the performances of these models are not so satisfactory due to the non-availability of annotated endoscopic databases. That is why, transfer learning approaches have been used to address the polyp classification problem. In the transfer learning approach, learned features of the existing models are used, and subsequently they are finetuned for polyp classification. However, the transfer learning method is not generally suitable for endoscopic image classification as real-time patient and device-specific endoscopic images are sometimes different from the images obtained by transfer learning. That is why, any model can give a very good performance during training, however, the same model may miserably fail during real-time polyp classification. Some of the existing deep learning-based polyp classification techniques are highlighted below.

Not many very promising methods are reported in the literature because of the unavailability of publicly available annotated data of colonoscopic polyps, and that is why it would not be possible to do the comparative analysis for different texture features. Mostly, handcrafted features are employed for polyp classification. Texture features are very important as endoscopists comprehensively analyze the texture patterns for finding dysplasia in polyps. The existing texture feature descriptors are

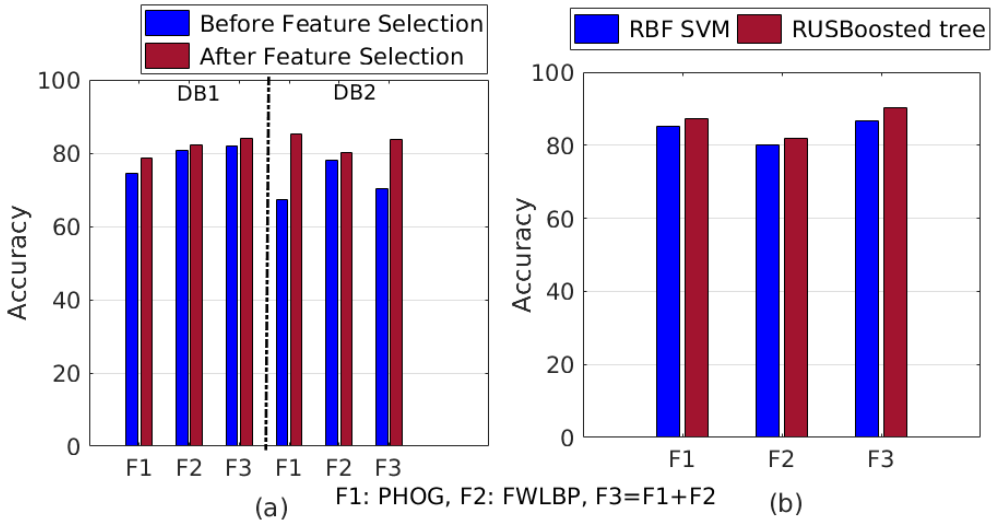


Figure 4.9: Performance analysis before and after feature selection (a) accuracy using RBF SVM for both the databases (b) comparision of accuracies for SVM and RUSBoosted tree classifiers for DB2.

also studied in our work. Song et al. [254] proposed locally encoded transform feature histogram (LETRIST) for texture classification. In [263], local grouped order pattern (LGOP) and nonlocal binary pattern (NLBP) operators were proposed for texture description. LGOP encodes the group-wise intensity order relationships and NLBP computes several anchors based on global image statistics and progressively encodes non-local intensity differences between the neighboring sampling points and anchors. Finally, LGOP and NLBP were combined to construct discriminative histogram features as LGONBP texture descriptor. Polyp classification results for some of these methods are shown in Table 4.5. In all of these methods, SVM is used for classification.

Table 4.5: Comparision of classification performance with different texture descriptors using SVM classifier

Method	Accuracy % (DB1)	Accuracy % (DB2)
LBP	70.5	72.2
HOG	64.5	68.9
GLCM	71.5	70.5
LETRIST [254]	78.5	80.2
LGONBP [263]	82.1	83.2
FWLBP [255]	82.2	80.2

The method proposed in [134] fuses three types of features –transfer learning features, fully trained CNNs features, and classical hand-crafted features. The combined features give a classification accuracy of 93.22%. The combined features used in their method are CNN-M MCN, VGG-VD16, and

Table 4.6: Comparative study of polyp classification accuracy between the baseline deep learning models and our proposed method

	Average accuracy
Proposed	**90.16**
VGG16	82
VGG16 fine-tuned	90
VGG19	70
VGG19 fine-tuned	89
MobileNet	90.11
ResNet50	68
ResNet50 fine-tuned	72
Inception-V3	90.02

blob shape adapted gradient using the local fractal dimension (BSAG-LFD). However, the extraction of altogether three different types of features increases the computational burden. Also, the model is tested on the dataset taken from high definition modality, where many image features like contrast and tone of the vascular tissues are clearly discriminable, which might not be available in the images captured by general endoscopic modalities. In [135], per-class data augmentation is adopted to tackle an unbalanced class distribution to improve classification and thereby improved classification accuracy. The classification accuracy of this method is 90.2%. Golhar et al. [136] proposed a semi-supervised learning approach for lesion classification in endoscopic images. They introduced a jigsaw puzzle solver into a semi-supervised learning model which uses an encoder to generate discriminative features. The classification accuracy of their method is 79.76%, which is tested on their own dataset. We earlier proposed a GAN-generated synthetic image augmentation scheme, followed by a conventional CNN for polyp classification [137]. Two databases were used to validate our method. One is a publicly available database and other one is our own generated database. The polyp ROIs were manually extracted from the colonoscopic video frames under the supervision of an expert in this domain. The classification accuracy of our method was 88.33%. Since a limited number of training samples are publicly available at present, deep learning approaches would not be suitable for handling this classification problem. A comparative study of polyp classification accuracy between our proposed method and the state-of-the-art deep learning-based methods is shown in Table 4.6.

To have a fair comparison of our method, we selected only those existing works which are related to the classification of the polyp. We selected the DB2-NBI database as it is widely used in the existing works. Also, only the classification of benign and malignant are considered for comparison. The comparative results with the state-of-the-art methods are listed in Table 4.7. It is observed from

Table 4.7: Comparison with the existing works

	Accuracy	Sensitivity	Specificity	Precision	Recall	F-Score
Proposed	90.16	92.50	85.71	92.50	92.50	92.50
2D texture+3D [128]	89.47	94.55	76.19	91.23	94.55	92.86
LoG [264]	84.21	90.91	66.67	87.72	90.91	89.29
Color GLCM [130]	64.47	74.55	38.10	75.93	74.55	75.23
BoW+SPM [131]	73.68	98.18	90.52	73.97	98.18	84.38

Table 4.7 that the proposed work outperforms the existing works in colonic polyp classification.

4.2.3 Conclusion

This work proposes using the shape and texture features of the polyps to classify the stages of dysplasia. For this, the local polyp shape features are extracted using PHOG, and the local texture features are extracted using FWLBP. The endoscopic video frames are prone to affine transformations. So, the characteristics of the polyp may be perceived differently for different ambient conditions. SPM and FD were incorporated with this framework to deal with this problem. A feature ranking algorithm based on fuzzy entropy is adopted for feature selection. The final assessment using different performance matrices establishes the efficacy of the proposed work. This method is also applied to our own dataset, which contains images from all the three modalities, viz., NBI, WL, and Dye. The consistency in performance demonstrates the robustness of our method. Thus, the selected features can serve the purpose of accurate polyp classification for different modalities. Deep learning-based polyp classification methods have a big challenge due to the lack of extensive and publicly available annotated databases. On the contrary, our method can be deployed to both offline and real-time colonoscopic polyp classification.

Though the proposed method achieves an acceptable accuracy, the learned texture features may not be sufficient to discriminate between polyp classes. The idea is to learn distributed embeddings representation of data points so that contextually similar data points are projected in the nearby region in the low dimensional vector space. In contrast, different data points are projected far away from each other. This technique, therefore, extract such features from the polyp classes that ensure inter-class separability. Thus, our next approach would investigate the effectiveness of the learned features via a similarity learning framework.

4.3 Feature Fusion-based Approach

During the colonoscopy, the captured video frames are stored in a computer system for further analysis in the future. However, real-time analysis of colonoscopy video frames can lead to better diagnosis and early treatment. Also, the decision support system must automatically detect any abnormality in the frames. Therefore, the first stage of our current work focuses on handling real-time data for automatic detection and localization of polyps in the colonoscopy frames. For this, a deep learning-based attention YOLOv4 model was proposed in Chapter 2. Following the identification of polyps, endoscopists split off the polyp areas and vividly access them for cancer diagnosis. An automated polyp classifier for two-class polyp classification, i.e., adenoma (malignant) and hyperplastic (benign), is provided in this work.

Hand-crafted feature learning approaches were used in the early research on automatic polyp categorization from colonoscopy frames [128,130,131,264,265]. The inconsistency of these approaches' performance in terms of repeatability is a drawback. Furthermore, the generalizability and robustness of these techniques cannot be guaranteed because pathological situations vary greatly even within the same modality's dataset. Also, a huge domain knowledge is required to characterize the discriminating features of the polyps. Deep learning-based techniques are better at handling such variances and give a high degree of generalisation. As a result, there has been an increase in interest in using such models in medical image and video processing *especially* in polyp classification [132–136]. However, one of the primary drawbacks of these methods is that they require a large quantity of labelled data during training in order to get relatively good classification performance. Large-scale polyp databases, on the other hand, are harder to achieve by. The wide range of imaging methods and processes, as well as privacy concerns and a lack of medical integration, may provide a number of obstacles in obtaining high-quality, large-scale polyp images. In light of these issues, we suggested a classification technique that does not necessitate the use of large amounts of labelled data. We'll illustrate how the non-linearity of a small, imbalanced dataset may be correctly described by the features learned via our proposed network. For classification of the localized polyps, we propose using the Triplet network architecture and its related triplet loss to learn non-linear representations between polyps. We show that the learned features may be used as a highly discriminative basis for machine learning models. We compare our findings to those of prior research and show that the features acquired by a Triple Network can characterize the non-linearity of a small dataset, making them acceptable

for use in a linear classifier. In addition, integrating deep and handcrafted features improves polyp classification efficiency. For deep features, we employed a triplet network based on siamese architecture and the handcrafted features were extracted using pyramid histogram of oriented gradient (PHOG). As discussed earlier, texture and shape information of polyps play a vital role while dysplasia grading by the endoscopists. In our proposed framework[2], the triplet network helps to learn distributed embedding by the notion of similarity and dissimilarity whereas the PHOG extracts the shape and texture information of the polyps [265]. The suggested classification approach is shown schematically in Figure 4.10. The rest of the work is organized as follows. Section 4.3.1 discusses the proposed methodology. The experimental results and conclusion are given in section 4.3.2 and section 4.3.3, respectively.

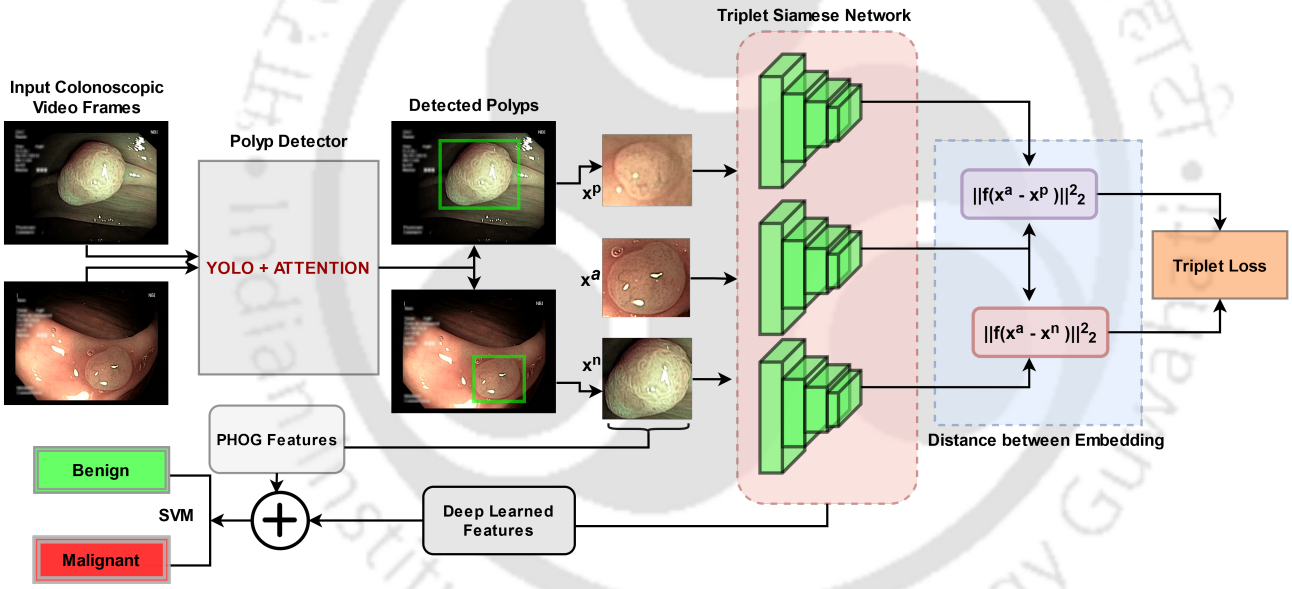


Figure 4.10: Proposed polyp classification approach.

4.3.1 Proposed method

The fusion of complementary features can boost the performance of a classifier. The proposed approach discusses the effect of combining embeddings and hand-crafted features on polyp classification.

4.3.1.1 PHOG

A polyp's geometry, texture, and colour provide enough information on its nature. The proposed approach uses a pyramid histogram of oriented gradient (PHOG) characteristics to define the geometry

[2]This work is part of the earlier discussed proposed approach, revision submitted to Scientific Reports, Nature (Refer *List of publications* page for details).

or morphology of a polyp. At each pyramid resolution level, the HOG vector is calculated (L). Finally, the PHOG descriptor is extracted by concatenating all of the HOG vectors. The PHOG descriptor's dimensionality for the full image is provided as: $\boldsymbol{K} * \sum_{l=0}^{\boldsymbol{L}} \boldsymbol{4^l}$. In this work, $\boldsymbol{K}$ and $\boldsymbol{L}$ values are taken as 8 and 4, respectively. The details of this feature extraction technique was discussed in the previous work.

4.3.1.2 Triplet network

The Siamese network [266] inspired Triplet network design consists of three identical sub-networks with common parameters. Each sub-network is taught to recognize embedded characteristics in three different samples, the anchor, positive, and negative samples, respectively. A triplet is made up of an anchor, a positive, and a negative sample. The anchor and positive samples were labeled as benign polyps, while the negative was labeled as malignant. Training a Siamese network using triplet loss requires two inputs. One pair has similar images from the same class, and the other should have images from different classes. Therefore, an anchor is selected from either class and accordingly, pairs are created. In this study, the anchor is selected from the benign class. The positive sample means the sample selected from the anchor class, whereas the negative class image belongs to a completely different class. The resulting embedding from the trained Triplet network is capable of separating the anchor from the negative class while maintaining the proximity between the anchor and the positive class. The triplet loss maximizes distance between opposite samples in a low dimensional space. The L2 distance between the anchor and the positive sample, as well as the anchor and the negative sample, are the network outputs. The cost function is computed using the triplet loss in Eq. 4.18, where $\boldsymbol{f_i^a}$ represents the anchor embedding, $\boldsymbol{f_i^p}$ represents the positive embedding, and $\boldsymbol{f_i^n}$ represents the negative embedding.

$$\boldsymbol{L = \max(0, ||f_i^a - f_i^p||_2^2 - ||f_i^a - f_i^n||_2^2 + \alpha)} \tag{4.18}$$

The value of $\boldsymbol{\alpha}$ was taken 0.5, and the dimensionality of the embedding was set as 256.

4.3.1.3 Training

The anchor and positive samples were labeled as benign polyps, while the negative was labeled as malignant. Three Triplet networks were trained using identical hyperparameters, with Adam as the preferred optimizer and learning rate 0.0001 as the hyperparameters. Each network was started using

ImageNet weights and trained from the ground up. For each of the triplet's images, our Siamese Network will generate embeddings. We achieved this by connecting a few Dense layers to a ResNet50 model that has been pre-trained on ImageNet. All of the model's layers' weights will be frozen until the layer conv5_block1_out. The last layers were fine-tuned during training. One Linear SVM for modality and fold was trained and assessed to examine the retrieved features.

4.3.2 Results and discussion

The proposed method is validated on the publicly available a labeled polyp dataset for colorectal polyps classification [128]. The dataset is available at url: `http://www.depeca.uah.es/colonoscopy_dataset/`. It contains video sequences using narrow-band imaging (NBI) and White light (WL) imaging. The dataset contains video sequences for 21, 15, and 40 hyperplastic (benign), serrated, and adenoma (malignant) polyps. Figure 4.11 shows some of the samples from both the classes of the dataset. The video sequences are converted to frames, and from each frames, the polyps are detected using the proposed YOLOv4 attention network. Subsequently, these polyps are fed to the triplet network for classification. In this work, only NBI image frames from hyperplastic and adenoma classes are considered. Three-fold cross-validation was employed as a validation method for our approach. Extracted features are analyzed using linear SVMs to classify polyps between benign and malignant. A classification accuracy of 90.16% is achieved. The embedded features of the image samples of the

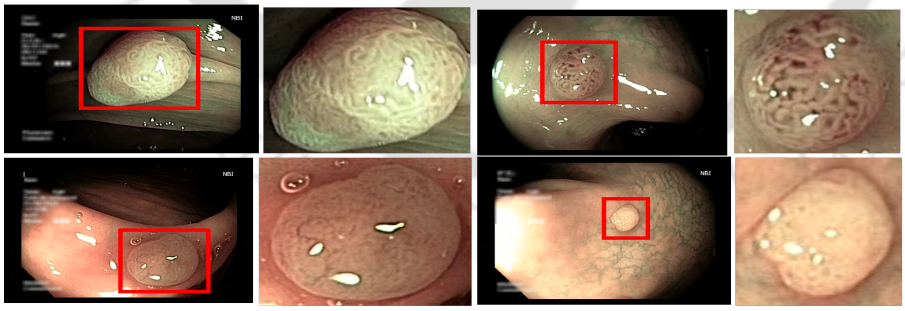
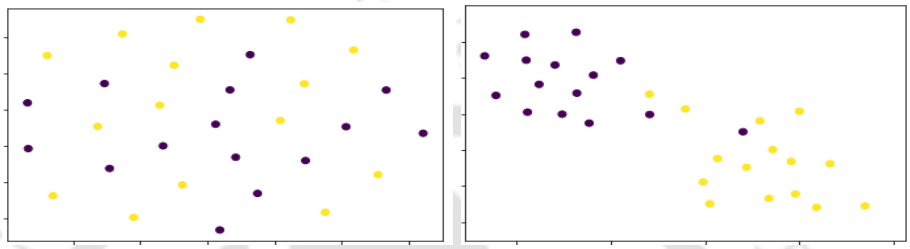


Figure 4.11: Sample frames from both the classes. First row samples are of malignant type and the bottom row images are of benign type. The polyps are detected by the YOLO-v4 attention model.

database are analyzed using t-SNE and are shown in Figure 4.12. Further, the PHOG features are also fused with the embedded features extracted from the triplet network to enhance the classification performances. The fusion of these features increases the dimensionality and non-linearity in the feature space. Therefore, an RBF kernel SVM was used for classification of the fused features. It was also varified from the experiments that the RBF SVM performs better as compared to other classi-

fiers. Table 4.8 shows the classification accuracies of some of the handcrafted-based methods on this publicly available colonoscopy dataset [128]. Similarly, Table 4.9 shows the classification accuracies on the same dataset using the transfer learning approaches. Finally, the results are compared with the state-of-the-art methods, and it is clearly seen that our method gives better performances in a limited data environment. The results are shown in Table 4.10.



(a) (b)

Figure 4.12: t-Distributed Stochastic Neighbor Embedding (t-SNE) is employed to decrease the dimensionality of the feature embedding into a 2D representation, . (a) initial features after the first training epoch, illustrating the embedding space mixture of classes with no distinction. (b): the final embedding once the model has converged to an optimum solution. Malignant polyps are shown by yellow dots, while benign polyps are represented by violet dots.

The comparison with other approaches has been extremely difficult for several reasons. Most of the methods available in the literature have used their own datasets which are private. Very few methods using this publicly available datasets are available. Therefore, Table 4.8 and Table 4.9 show the performances of various feature learning approaches on the polyp classification on the publicly available dataset. Methods shown in Table 4.10 have also used the same dataset for a fair comparision of perfromance with our proposed method.

Table 4.8: Comparision of classification performance with differfent texture descriptors using SVM classifier.

Method	Accuracy
LBP	72.29
HOG	68.93
GLCM	70.52
Curvelet features [131]	70.90
LETRIST [254]	80.27
LGONBP [263]	83.20
FWLBP [255]	80.25
PHOG	85.20
PHOG+FWLBP [265]	90.16

Table 4.9: Comparison of classification accuracy between the baseline deep learning models and our method.

Method	Accuracy
Proposed	**96.66**
VGG16 [267]	82
VGG16 fine-tuned	90
VGG19 [267]	70
VGG19 fine-tuned	89
MobileNet [268]	90.11
ResNet50 [142]	68
ResNet50 fine-tuned	72
Inception v3 [269]	90.02

Table 4.10: Comparison with the existing works. Acc.—Accuracy, Sen.—Sensitivity, Spec.—Specificity, Pre.—Precision, Rec.—Recall.

	Acc.	**Sen.**	**Spec.**	**Pre.**	**Rec.**	**F1-Score**
Proposed	**96.66**	93.33	93.75	93.33	100.00	96.54
2D Texture+3D Features [128]	89.47	94.55	76.19	91.23	94.55	92.86
LoG [264]	84.21	90.91	66.67	87.72	90.91	89.29
Color GLCM [130]	64.47	74.55	38.10	75.93	74.55	75.23
BoW+SPM [131]	73.68	98.18	90.52	73.97	98.18	84.38
Triplet network [270]	90.16	92.50	85.71	92.50	98.25	92.50
PHOG+FWLBP [265]	90.16	92.50	85.71	92.50	92.50	92.50
NSCT+GFD [271]	95.72	95.31	95.00	93.22	92.15	90.45

The two databases used for validating our proposed method for polyp detection suggest that the algorithm could detect various polyp structures quite accurately. It is crucial in clinical practice as early detection helps better prognosis and clinical management, leading to a higher survival rate. Following detection, the localized polyps are classified, which is crucial for early diagnosis and better prognosis. In this view, a two-class polyp classification system is proposed in this study. The polyp classifier system must provide good classification performances. However, the traditional learning-based approaches generally provide low classification performances [40-44]. Our earlier work in polyp classification achieved a classification accuracy and F1-score of 95.72% and 90.45%, respectively [271]. In this work, shape feature was extracted by the Non-subsampled contourlet transform (NSCT) and shape information was extracted by Gaussian Fourier Descriptor (GFD). From, the results, it was inferred that texture and shape can be used to characterize polyp features.

Typically, deep-learning-based techniques require good quality annotated images for training the models [45-49]. Also, the training of such models needs balanced data from each class. However, having a similar number of images from all categories of polyps is quite challenging in colonoscopy procedures. Generally, cancerous polyps are found in small numbers compared to other varieties. Also, different classes of polyp images have subtle differences in features. Therefore, simply learning features from each polyp class may not help better discrimination. Thus, similarity learning approaches have been adopted to extract non-linear feature representations of the polyp classes [50]. The features learned from the proposed Triple network of the siamese architecture can describe the non-linearity of a relatively small and imbalanced dataset. Additionally, local polyp features extracted by PHOG are fused with the embeddings, resulting in improved classification results. The effectiveness of our strategy in a limited data environment is demonstrated by its classification performance on a relatively

small dataset.

We aim to work on these issues in the future, especially the impact of noise on classification performances. Semisupervised learning-based (SSL) approaches are good at learning generic features from the unlabeled data and discriminative features from the labeled data and thus help in better polyp characterization. It can leverage the requirements of large annotated datasets by the supervised deep models. As the colonoscopy procedure has limited annotated polyp images and relatively higher unlabeled images, approaches based on SSL can be adopted. Further, sub-grading of dysplasia in polyps can be done, which could also allow practitioners to comprehend pathological situations in a better way.

4.3.3 Conclusion

This approach presents a framework for the analysis of colonic polyps using colonoscopy video frames. A deep attention based YOLOv4 network is proposed to detect polyps, followed by their classification. The performance of the suggested algorithm outperforms state-of-the-art approaches by a significant margin. We propose a triplet network based on siamese architecture, followed by SVM, to classify the polyps. Additionally, local polyp features are extracted and fused with deep features, resulting in improved classification results. The effectiveness of our strategy in a limited data environment is demonstrated by its classification performance on a relatively small dataset. Further, grading of dysplasia in polyps could also allow practitioners better comprehend pathological situations.

Histopathological images of the polyps are generally studied in clinical setup to detect CRC. These images provide a detailed insight into polyp characteristics for dysplasia grading. Due to the lack of a large annotated dataset in this domain, an SSL-based approach is proposed for this work. The proposed GAN framework can learn useful features from the unlabeled histopathological images in the unsupervised mode, whereas it can classify the polyps in the supervised mode. Therefore, this framework can be suited for a limited data environment and is generally encountered in polyp diagnosis.

4.4 A Semisupervised GAN for Classification

The need for histopathological images for better clinical interpretation is on the rise [272]. Gastrointestinal histopathologists analyze the tissue samples acquired during colonoscopy. Histological specimens of colorectal polyps are pertinent for the experts in cancer diagnosis. The high resolution of such images enables complete characterization of colorectal polyps for early and promptly diagnosing an invasive carcinoma.

Deep learning models have outperformed humans in the computer vision area during the last decade on tasks including image categorization and object recognition, in some cases surpassing human performance. These models have surpassed traditional image processing techniques in various medical imaging domains, such as radiology, histology, retinopathy, and mammography. Most of these models are trained in a supervised manner to attain optimal performance, requiring enormous amounts of expertly annotated medical data. Compiling annotated data in the medical imaging field is exceptionally time-intensive, costly, laden with privacy problems, and limited by the availability of expert annotators. In contrast, unsupervised methods have shown that meaningful representations can be extracted from unlabeled data, which is often plentiful. In this work, we leverage the advantages of both labeled and unlabeled data using the semi-supervised learning (SSL) paradigm to improve the performance of colonoscopy lesion classification. SSL is a new field of study that tries to learn a supervised goal while enhancing the encoded features via an unsupervised task. Recent research has demonstrated significant improvements over simply supervised training, particularly with modest amounts of labeled data [273, 274].

Other researchers worked on histopathological images for colorectal polyp classification. Korbar et al. [141] offered a patch-based framework for classifying different types of colorectal polyps from whole-slide images, which was deployed using a ResNet architecture [142]. Wei et al. [143] created a hierarchical classification approach to match the nature of the classification problem for the deep learning model to infer the overall diagnosis of a whole-slide image. Using a sliding window algorithm, each slide was first split down into several patches, and the deep ResNet-like neural network then classified each patch. Song et al. [144] proposed a patch-based fully convolutional technique for adenomas classification and grading, with a significant emphasis on model interpretability. They also show how different patch sizes should be used to classify and grade adenomas.

However, the requirement of a considerable amount of labeled data during training for reasonably

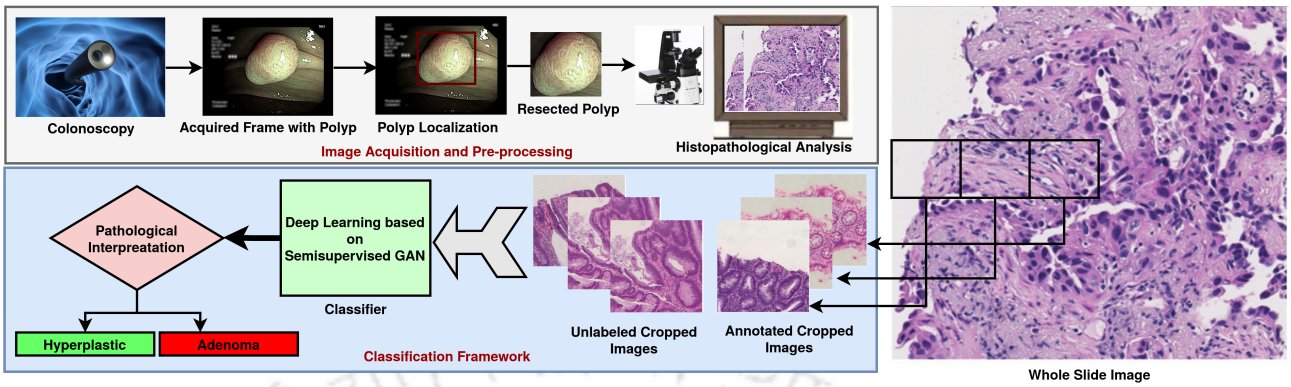


Figure 4.13: Flow diagram of a typical polyp classification system using our proposed deep learning-based classifier.

excellent classification performance is one of the significant limitations of these approaches. Obtaining large-scale polyp datasets, on the other hand, poses significant hurdles. The quality of the images obtained is suffered by many adversaries like noise, blur, specularity, stool, blood, bubbles, and other factors. The huge variances in imaging modalities and procedures, privacy issues, and lack of medical integration may pose many challenges in acquiring high-quality and large-scale polyp images. Furthermore, the acquired datasets are typically unbalanced, as some diseases are more common than others. Secondly, the burden on the experts, inter and intra evaluator variabilities are common in these procedures, necessitating additional protocols to resolve them.

This work proposes an SSL approach based on a generative adversarial network (GAN) in a limited labeled data scenario for polyp classification. The proposed GAN is extended to the semi-supervised context to create a data-efficient classifier by forcing the discriminator network of the model to output the class labels. The flow diagram of a typical polyp classification procedure using our proposed classifier is represented schematically in Figure 4.13. The proposed methodology is described in section 4.4.1. section 4.4.2 and section 4.4.3 provide experimental results and conclusion, respectively.

4.4.1 Proposed method

The proposed method is based on an SSL approach. A GAN-based framework is proposed to classify the histopathology polyp images.

4.4.1.1 Classical GAN

GANs are a class of game theory-based approaches for learning generative models in an adversarial manner [275]. It consists of two neural networks, i.e., the generator $\boldsymbol{G}$ and the discriminator $\boldsymbol{D}$. These

are used to train a generator network $G(z;\theta^{(G)})$ that generates samples from a data distribution, $p_{data}(x)$, by transforming noise vectors z as $x = G(z;\theta^{(G)})$. The generated images x_g and the real images x_r are classified as fake and real, respectively, by D. D has been trained to maximize the likelihood of correctly labeling both actual and fake images. G has been taught to deceive the discriminator into thinking the generated samples are authentic. The generator's ability to generate increasingly realistic images improves during training, while the discriminator's ability to distinguish actual from synthesized images improves. The discriminator D is trained to maximize log(D(x)), whereas the generator G is trained to minimize $\log(1 - D(G(z)))$. As a result, the optimization's goal is to solve the minimax issue in the following way:

$$\min_G \max_D F(D,G) = E_{x\sim p(x)}[\log(D(x))] + E_{z\sim p(z)}[\log(1 - D(G(z)))] \tag{4.19}$$

Because of this adversarial learning, the generator is capable of producing realistic images. As a result, GANs have been deployed in a variety of applications, including image synthesis [275], inpainting, denoising [276], style transfer [277], and image superresolution [278–280]. Recently, GAN has been used in a number of medical imaging analyses and applications [276, 281, 282]. While these GAN architectures make use of the generator output, Odena [283] modified the discriminator and made it do multiclass classification. He was able to make the GANs act like a semisupervised framework. The new architecture resulted in a classifier that was more data-efficient. The MNIST and CIFAR-10 datasets were used to achieve preliminary test results [284]. The improved GAN's data-efficient feature makes it particularly well suited to categorizing clinical, histopathological images, which are not readily available.

4.4.1.2 Proposed classification framework using GAN

The suggested GAN architecture for histopathological image classification is depicted in Figure 4.14 as a block diagram. The semisupervised GAN is a variation of the traditional GAN network for predictive modeling with a small number of labeled colon histopathological images $\{X_l, y_l\}$ from N classes and a large number of unlabeled images X_u with a distribution comparable to the labeled samples [283]. It consists of a modified discriminator (D_m) and a modified generator (G_m). In the proposed framework, the D_m has two modes of operation: supervised and unsupervised. The discriminator is trained in the supervised mode to predict the class labels for real histopathological images. The discriminator

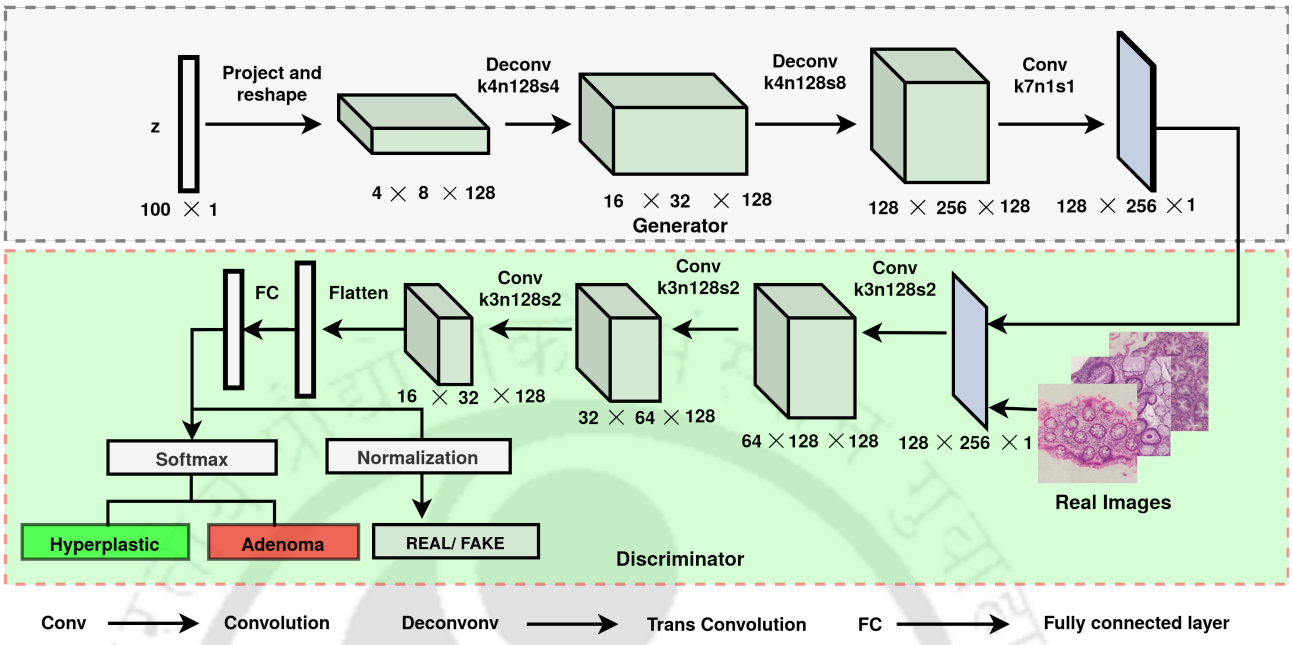


Figure 4.14: Block diagram of the semisupervised GAN for the histopathological image classification.

performs the same goal as the standard GAN framework in unsupervised mode: classifying images as real or fake. The discriminator can learn useful features from the unlabeled histopathological images by training in the unsupervised mode.

A shared discriminator architecture, as shown in Figure 4.14, is used to design $\boldsymbol{D_m}$ for operation in both modes. The architecture has two separate models for supervised and unsupervised modes, both of which use the same network parameters. The output units of $\boldsymbol{D_m}$ in the supervised mode correspond to the Softmax output, i.e., [$\mathbf{class_1}$, $\mathbf{class_2}$, ..., $\mathbf{class_N}$]. These $\boldsymbol{N}$ classes represent the class probabilities for the various CRC diseases. In the unsupervised mode, the outputs before Softmax activation are normalised as $\boldsymbol{N(x)} = \frac{\boldsymbol{J(x)}}{\boldsymbol{J(x)+1}}$, $\boldsymbol{J(x)} = \sum_{i=1}^{N} \mathbf{exp}\{\boldsymbol{l_i(x)}\}$, where $\boldsymbol{l}$ is the feature vector for input image $\boldsymbol{x}$ before Softmax activation. For real unlabeled and fake synthesized images, the normalized outputs produce the probability of real/fake ($\mathbf{class_{N+1}}$). The method proposed in [284] provided the inspiration for the discriminator losses for both modes, which are expressed as

$$\boldsymbol{L}_{\mathbf{sup}} = -\boldsymbol{E}_{\boldsymbol{x},\boldsymbol{y}\sim\boldsymbol{p}_{\boldsymbol{data}}(\boldsymbol{x}_l,\boldsymbol{y}_l)} \mathbf{log}\ \boldsymbol{p}_{\boldsymbol{model}}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{y} < \boldsymbol{K}+1) \tag{4.20}$$

$$\boldsymbol{L}_{\mathbf{unsup}} = -\{\boldsymbol{E}_{\boldsymbol{x}\sim\boldsymbol{p}_{\boldsymbol{data}}(\boldsymbol{x}_u)} \mathbf{log}\ [1 - \boldsymbol{p}_{\boldsymbol{model}}(\boldsymbol{y} = \boldsymbol{K}+1|\boldsymbol{x})] + \boldsymbol{E}_{\boldsymbol{x}\sim\boldsymbol{G}} \mathbf{log}\ [\boldsymbol{p}_{\boldsymbol{model}}(\boldsymbol{y} = \boldsymbol{K}+1|\boldsymbol{x})]\} \tag{4.21}$$

where, $\boldsymbol{L_{sup}}$ and $\boldsymbol{L_{unsup}}$ are the supervised and unsupervised discriminator losses, respectively. $\boldsymbol{p_{data}(x|y)}$ is the probability distribution of $\boldsymbol{D_m}$, $\boldsymbol{p_{data}(x_l, y_l)}$, $\boldsymbol{p_{data}(x_u)}$, and $\boldsymbol{G}$ are the probability distribution of the $\mathbf{X_l}, \boldsymbol{y_l}$, $\mathbf{X_u}$, and the generated fake images by $\boldsymbol{G_m}$, respectively. $\boldsymbol{E\{\cdot\}}$ is the expectation operation.

where the whole cross-entropy loss has been decomposed into our normal supervised loss function $\boldsymbol{L_{sup}}$ (the negative log probability of the label given that the data is real) and an unsupervised loss function $\boldsymbol{L_{unsup}}$ which is, in fact, the standard GAN game-value as becomes evident when we substitute $\boldsymbol{D(x) = 1 - p_{model}(y = K + 1|x)}$ into the expression as given in Eq. (4.21):

$$\boldsymbol{L_{unsup} = -\{E_{x\sim p_{data}(x_u)} + E_{z\sim p(z)}[\log(1 - D(G(z))]\}} \tag{4.22}$$

The generator loss is given by

$$\boldsymbol{L_g = E_{x\sim G}\log\big[1 - p_{data}(y = K + 1|(x)\big]} \tag{4.23}$$

The discriminator and generator losses are minimized in training. The sparse and binary categorical cross-entropy loss functions are used to achieve the supervised and unsupervised losses. The Algorithm summarises the training process for the proposed framework. The final findings are unaffected by modifying the update sequence of $\boldsymbol{D_m}$'s supervised and unsupervised losses in the method. The supervised $\boldsymbol{D_m}$ network is used to classify polyps at the end of the training phase, while $\boldsymbol{G_m}$ is discarded because it was only used to help $\boldsymbol{D_m}$ during the training process.

Algorithm 1 Training Algoritm for the Proposed Framework

Input: $\boldsymbol{I}$: number of total iterations, set of total labeled and unlabeled histopathological images $\{\boldsymbol{X_l}, \boldsymbol{y_l}\}$ and $\{\boldsymbol{X_u}\}$

1: **for** $\boldsymbol{i = 1}$ to $\boldsymbol{I}$ **do**

2: Sample $\boldsymbol{k}$ samples from the noise prior $\boldsymbol{p_z(z)}$
Sample $\boldsymbol{k}$ histopathological images each from the $\{\boldsymbol{X_l}, \boldsymbol{y_l}\}$ and $\{\boldsymbol{X_u}\}$
Generate fake image samples: $\boldsymbol{X_f}$=$\boldsymbol{G_m(z)}$, $\boldsymbol{X_f \in G}$
Perform gradient descent on the parameters of $\boldsymbol{D_m}$ using $\{\boldsymbol{X_l}, \boldsymbol{y_l}\}$ by computing the gradient as follows:
$\boldsymbol{\Delta_{\theta D_m}} \frac{\mathbf{1}}{\boldsymbol{k}} \sum_{\boldsymbol{j=1}}^{\boldsymbol{k}} \boldsymbol{L_{sup}}$ (Loss computed using Eq. 2)
Perform gradient descent on the parameters of $\boldsymbol{D_m}$ using $\{\boldsymbol{X_u}\}$ and $\{\boldsymbol{X_f}\}$ by computing the gradient as follows:
$\boldsymbol{\Delta_{\theta D_m}} \frac{\mathbf{1}}{\boldsymbol{2k}} \sum_{\boldsymbol{j=1}}^{\boldsymbol{2k}} \boldsymbol{L_{unsup}}$ (Loss computed using Eq. 3)
Perform gradient descent on the parameters of $\boldsymbol{G_m}$ by computing the gradient as:
$\boldsymbol{\Delta_{\theta G_m}} \frac{\mathbf{1}}{\boldsymbol{k}} \sum_{\boldsymbol{j=1}}^{\boldsymbol{k}} \boldsymbol{L_g}$ (Loss computed using Eq. 5)

3: **end for**

4.4.2 Results and discussion

4.4.2.1 Dataset and experimental setup

The proposed method is validated on the publicly available UniToPatho, a labeled histopathological dataset for colorectal polyps classification [26]. The dataset is available at the following url: `https://ieee-dataport.org/open-access/unitopatho`. UniToPatho is a collection of annotated high-resolution Hematoxylin and Eosin, H&E-stained images, containing various colorectal polyp histology samples acquired from cancer screening patients. According to UniTo pathologists' judgment, the dataset is a collection of the most relevant patch images derived from 292 whole-slide images. The slides are scanned at 20× magnification (0:4415 μm/px) using a Hamamatsu Nanozoomer S210 scanner, as shown in Figure 4.15. Each slide belongs to a separate patient and is annotated by UniTo pathologists into the following six categories: 1) NORM – Normal tissue, 2) HP – Hyperplastic Polyp, 3) TA.HG - Tubular Adenoma, High-Grade dysplasia 4) TA.LG – Tubular Adenoma, Low-Grade dysplasia 5) TVA.HG – Tubulo-Villous Adenoma, High-Grade dysplasia 6) TVA.LG – Tubulo-Villous Adenoma, Low-Grade dysplasia. Figure 4.15 shows some of the high resolution histopathological image samples of

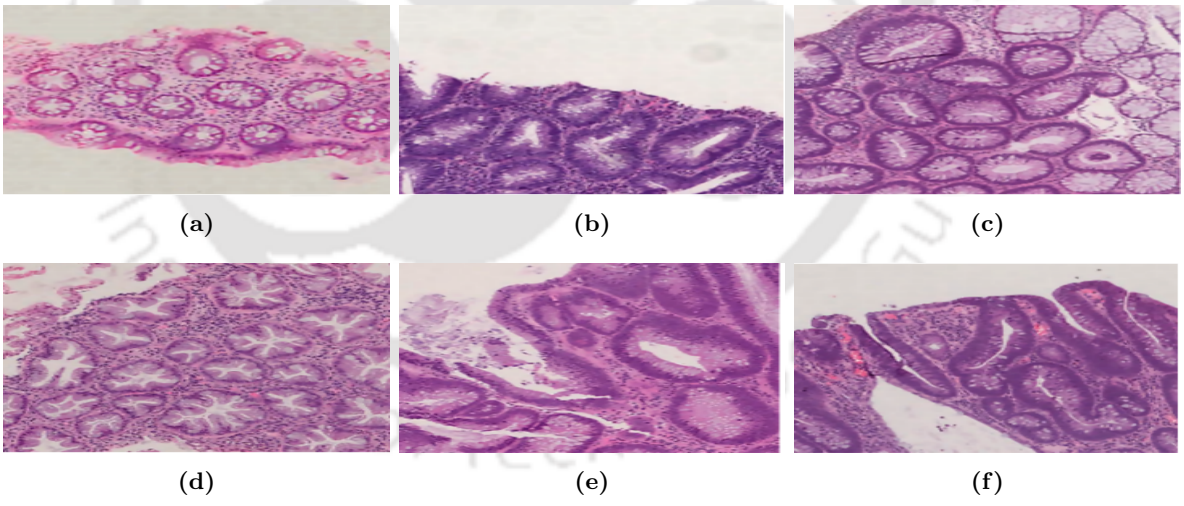


Figure 4.15: Some histopathological samples of polyps belonging to different classes; (a) normal tissue (b) tubular adenoma, high-grade dysplasia (c) tubular adenoma, low-grade dysplasia (d) hyperplastic (e) tubulo-villous adenoma, low-grade dysplasia (f) tubulo-villous adenoma, high-grade dysplasia. The image samples are taken from the UniToPatho dataset.

polyp tissues belonging to different classes. The details of the dataset is given in Table 4.11. Adenomas are more likely to evolve into invasive carcinomas than hyperplastic polyps, which normally have little malignant potential [285]. Therefore, accurate classification of adenoma polyps from the hyperplastic polyps is vital. In this work, we consider only two class classification viz. hyperplastic (HP) and

adenoma (AD) for their clinical relevance.

The slides are divided into a train set and a test set in a 70:30 ratio, resulting in 204 slides used for training and 88 slides used for testing. Non-overlapping square patches at various scales from each slide are cropped. We provide a total of 9536 patches available to the public, including 8669 extracted at $\boldsymbol{\sigma}$= 800 (1812×1812 pixels patches) and 867 extracted at $\boldsymbol{\sigma}$ = 7000 (15,855 ×15;855 pixels patches). In this work, the patches cropped at $\boldsymbol{\sigma}$ **= 800** are considered for training and testing the model as this resolution has more image samples. Combining samples from both classes in training causes bias in the testing phase. From the experimental results, it has been observed that training the model with image samples of both resolutions learns only resolution information. From the original paper [26], it is evident that hyperplastic polyps are best classified at a finer 800m scale. On the other hand, tubular Adenomas (TA) and Tubulo-Villous Adenomas (TVA) are best categorized at a coarser 7000m scale. Also, it is established that benign polyps are best differentiated by looking at smaller-scale details such as gland edges [286]. In contrast, the adenomatous polyp is best distinguished by examining large-scale macro features such as complete gland shapes [287]. In this work, the test images from both the polyp classes are derived from the original annotated dataset of $\boldsymbol{\sigma}$ **= 800** resolution. It contains 2071 adenomatous and 102 hyperplastic images. The number of training samples is taken with different percentages of the total samples. The samples are annotated with labeled and unlabeled with different ratios to validate the model's efficiency. The details of the dataset are provided in Table 4.11.

Table 4.11: Summary of the dataset; Whole image slides (top) and the two patch scales (bottom).

	HP	NORM	TA.HG	TA.LG	TVA.HG	TVA.LG	**Total**
Slides	41	21	26	146	20	38	292
$\boldsymbol{\sigma}$ = 7000	59	74	98	411	93	132	867
$\boldsymbol{\sigma}$ = 800	545	950	454	3618	916	2186	8699
Total	604	1024	552	4029	1009	2318	9536

Figure 4.14 shows the number of filters, kernel sizes, and strides in the convolution and deconvolution layers. The leakyReLU activation functions with a negative slope of 0.2 are followed by the learnable layers of the generator (excluding the last convolution layer). At the generator, deconvolution layers are followed by batch normalisation layers. The last convolution unit contains Tanh activation. Convolution layers are followed by leakyReLU activation with a negative slope of 0.2 and

batch normalization layers on the discriminator side. Adding a mini-batch discrimination layer [284] to the discriminator network prevents mode collapse in the GAN. To avoid overfitting, an optimal drop-out factor of 70% is applied to all but the first convolution layer of the discriminator. The model employs an Adam optimizer with a learning rate and $\boldsymbol{\beta_1}$ of $\mathbf{2 \times 10^{-4}}$ and 0.5, respectively. The model was trained on minibatches of size 32 for 50 epochs. The experiments were performed on an NVIDIA TITAN Xp GPU.

4.4.2.2 Classification performance

Table 4.12 shows the different combinations of numbers of labeled and unlabeled training images and their performances on classification. The best-case scenario is selected from the patch-wise classification result, and the model is said to be optimal. From the result shown in Table 4.12, it can be inferred that the performance with the patch level classification accuracy is acceptable even with small labeled images. However, the patch-level classification accuracy is less because of contextual and spatial information loss. Also, the patches which are taken from the boundary of a whole slide merely contain any clinical information. Therefore, all the patches of the entire slide may not have sufficient clinical information for a reliable prediction. Therefore, a majority voting scheme is adopted in our work to grade the whole slide image. In this scheme, patches corresponding to a whole slide are first tested with the model. If the percentage of the patches of the same slide is predicted to belong to a particular class based on a set threshold, then the whole slide is annotated with the same grade of dysplasia. e.g., For 25% voting, if 25% of the test images are classified to a particular grade, the whole slide is said to belong to this grade of dysplasia. The 25% voting gives an overall accuracy (OA) of 87.50% compared to 76.25% obtained for 50% voting. In cancer diagnosis, a soft threshold must be adopted to increase the sensitivity of the classifier. Therefore, a 25% voting scheme is reasonable to use for a reliable prediction. Experimental results show that our proposed model can provide a better dysplasia grading classifier with minimal data.

From Table 4.13, it can be seen that with 25% voting, the TV-HG and TVA-HG have a classification accuracy of 100%. TVA is a more severe type of polyps than the TA. Our proposed method could classify the TVA-LG, TVA-HG, and TA-HG with very high confidence. Classification accuracy of 82.60% is obtained for TA-LG. Similarly, 90% accuracy is achieved for the hyperplastic class classification with this voting scheme. The low-grade TA patches are misclassified compared to other grades. The TA-LG is the first stage of cancer; therefore, the dominant features of the polyps of this

grade may not be very discriminating from the features of the hyperplastic polyps.

Table 4.12: Patchwise classification performance for different training protocols. L —Labeled, Un —Unlabeled, CW Sen. —Classwise sensitivity, CW Acc. —Classwise accuracy, Over Sen. —Overall sensitivity, Over. Acc. —Overall accuaracy.

	Labelled		Unlabelled		CW Sen.		CW Acc.		Over. Sen.	Over. Acc.
	AD	HP	AD	HP	AD	HP	AD	HP	OSe	OA
Less AD in L	70	223	222	223	53.5	85.29	54.99	54.99	69.4	54.99
Less (AD in L + HP Un)	70	223	222	70	49.01	94.12	51.13	51.13	71.56	51.13
Best case	169	274	274	169	71.61	58.82	71.01	71.01	65.22	71.01
HG L + HG Un	222	222	221	221	76.05	64.71	75.51	75.51	70.38	75.52
LG L + LG Un	222	222	221	221	84.5	37.25	82.28	82.28	60.88	82.28

Table 4.13: Classification performance on the Whole slide histopathology images based on majority voting.

	Adenoma				Hyperplastic
Grade and Type	TA-LG	TA-HG	TVA-LG	TVA-HG	
Total number of slides	46	9	8	7	10
Slides identified with 100% voting	13	3	1	2	6
Slides identified with above 75% voting	20	3	4	6	8
Slides identified with above 50% voting	32	7	7	6	9
Slides identified with above 25% voting	38	9	7	7	9

Table 4.14: Comparision of classification performances with the baseline methods.

Methods	Overall Accuracy (OA)
Baseline [142]	0.46
Multi-resolution Ensemble [26]	0.67
Proposed (25%) voting	**0.8750**
Proposed (50%) voting	**0.7625**

Table 4.14 shows the comparision of classification performances with the baseline methods. Figure 4.16 shows some of the misclassified histopathological samples predicted by our algorithm. Hyperplastic polyp nuclei are often tiny, oval, and scattered, whereas adenomatous polyp nuclei are dark and elongated [288]. The absence of prominent nuclei could explain all three misclassified adenomatous images. It can also be illustrated that the class of low-grade tubular adenoma and hyperplastic has a lot of uncertainty due to their similar expressions. Therefore, we would like to extract features from the nuclei of the histopathological images for classification in the future. Similarly, some images are blurred and the nuclei are not visible clearly. Therefore, these images might have been classified wrongly.

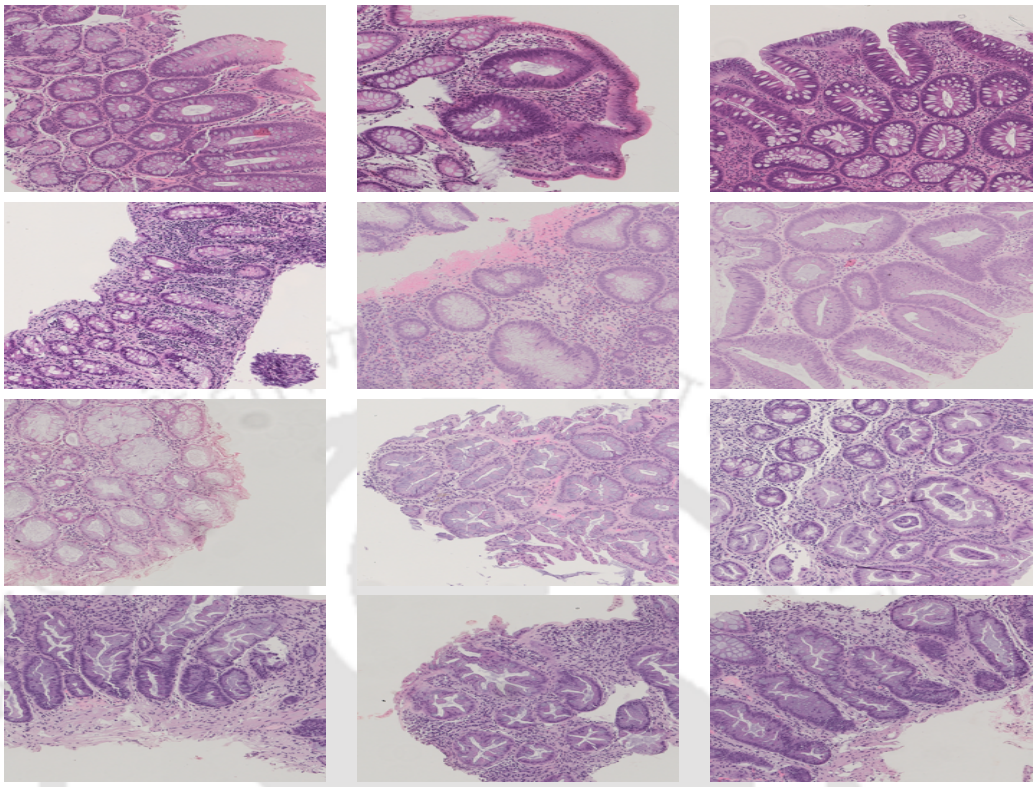


Figure 4.16: First row: correctly identified adenoma patches, second row: adenoma patches misclassified as hyperplastic, third row: correctly classified hyperplastic patches, and fourth row: hyperplastic patches incorrectly classified as adenoma.

4.4.3 Conclusion

In this work, we introduced and assessed a new automated GAN-based classifier for detecting carcinoma in polyps from histopathological images. The training of the model is done in a semisupervised manner for better feature representation of the colonic polyps. The suggested methodology has a substantial advantage in terms of generalizability, even under very restricted labeled data situations. It eliminates the time and expense of large-scale image acquisition and image annotation. Its superior classification performance makes it suitable for prescreening and automated diagnosis of colorectal cancer in hospitals. In future, we would like to study more on other techniques in the SSL paradigm for better characterization of polyps. We would also like to investigate how the SSL gains change as the size of the labeled and unlabeled datasets grows, i.e., to develop an interpretable model. SSL's domain evaluation enhancement could be extended by comparing not just imaging modalities but also different endoscopes with different resolutions, lighting settings, and frame rates.

4.5 Summary

This chapter proposes three methods for the classification of polyps. Geometry (shape and size), the surface of the polyp (texture), color (blood traces), boundary (smooth or wavy), etc., are analyzed to classify polyps into different grades of carcinoma. The first two methods are proposed to develop a two-class polyp classifier, i.e., adenoma (cancer) and non-adenoma (non-cancer). The shape and texture features are effectively extracted to discriminate the polyps types in the first approach. The shape of adenomatous polyps is wavy, and the region is not smooth compared to the non-adenomatous polyps. Therefore, a region-based shape detector, PHOG, is proposed to extract the feature. Similarly, the adenomatous polyps are highly textured, contrary to the normal polyps. The proposed FWLBP extracts features and is robust to geometric and photometric transformations, which the colonoscopy images are susceptible to. However, sometimes the texture features of polyps from both the classes are indistinctive. To extract features from the small and imbalanced dataset effectively, we adopted a similarity learning approach based on the Triplet network in our second approach. Further, fusing the shape features with the extracted embeddings from the Triplet network enhances the classification performance. Sometimes the polyp images may not be sufficient to extract useful features for their grading and sub-gradings. Therefore, histopathology images of such polyps are analyzed. In this view, our third approach is based on the GAN framework in an SSL paradigm. This approach is well suited to limited data scenarios and provides a good performance. In the future, we would like to investigate more on these images for developing polyp classification systems. Further, combining features from both the colonoscopy and histopathology images may be utilized to characterize polyps effectively.

5

Summary and Conclusions

Contents

Objective

This chapter summarizes the contributions of this thesis to the development of an automated polyp detection and classification system under various scenarios. The contributions of the work are mentioned. Future research directions are also outlined.

5.1 Summary

This thesis work aims to develop methods for automated polyp analysis using colonoscopy videos under varied scenarios. The work is broadly divided into three sub-works; polyp detection and localization, polyp segmentation, and polyp classification.

(i) **Polyp detection:** For any practical application of an automated polyp detector, the acquired images may be affected by background noise, non-uniform illuminations, photometric and geometric transformations, ghost colors, and normal tissues mimicking polyp, etc. In such a scenario, a better system can be developed by selecting the discriminative and dominant cues of the polyp. A polyp detection method is proposed by using a saliency map in a tracking framework based on particle filtering by localizing polyps in colonoscopy videos. This work shows how the saliency map can be used to choose particle weights and makes the tracking process faster. The inherent features of a polyp are used for the generation of such maps. Texture and color are two discriminative features for polyp non-polyp regions. The shape of the polyp is used for refinement of the ROI using an AC model. It thus helps in discarding the specular regions and converges towards better localization of polyps. The localized polyps are further segmented by AC. The experimental results show that our method is competitive with the state-of-the-art techniques in polyp segmentation. An experimental study on the CVC-ClinicDB database and the ETIS-Larib database shows that our method can be used effectively in polyp localization. Our approach is used for offline endoscopic video processing. The proposed method sometimes fails to localize small patchy polyps and polyps in highly over-exposed regions. The second work in this view presents a deep attention-based YOLOv4 framework to detect polyps by bounding box localization. The attention module in the YOLOv4 encapsulates spatial and contextual information of the polyp ROIs effectively. The attention module selectively accentuates the polyp ROIs by extracting local and global information. We used spatial and channel attention at different scales in the efficient YOLOv4 architecture. The performance of the suggested algorithm outperforms

state-of-the-art approaches by a significant margin. The consistency of results across datasets also demonstrates the generalizability and robustness of our method. The proposed method can do real-time polyp analysis. Following polyp detection, the polyps are classified, which is crucial for CRC diagnosis. The classification methodology is discussed in the corresponding chapter.

(ii) **Polyp segmentation:** Segmentation of polyps is more refined localization, i.e., perfect delineation of the polyp boundary. It can be formulated as another way of polyp detection by precisely detecting polyp boundaries. It is required for polyp resection and 3-D analysis of polyps. Also, the fewer the non-polyp pixels present in the localized polyp ROIs, the better the features representation of the polyp regions. However, the difficulty in achieving perfect polyp delineation is because of complex polyp backgrounds, indiscriminative polyp boundaries, etc. Big and convex polyps segmentation is easier than patchy and serrated polyps.

During the colonoscopy, uninformative and lousy images are often captured that contain no diagnostic information. Therefore, it is better to discard them before analyzing the suitable frames. The first method, therefore, does pre-process before segmenting the polyps. In this approach, the non-informative frames are discarded, and clinically significant polyp frames are retained. The essential frames are selected from colonoscopy videos with the help of depth information and the proposed three criteria selection strategy. Our proposed method determines depth maps using a zero-shot learning approach. The zero-shot learning method performs well on previously unseen classes like endoscopic images. Through this, we extended MDE to in-vivo images, which would be helpful to analyze colonoscopy images. We used texture, edge information from the depth maps, and the number of key points for selecting key-frames. It is to be noted that good frame polyps have vital clinical features compared to bad frames. Experimental results show the efficacy of the proposed method in selecting key-frames from endoscopic videos and subsequent segmentation of detected polyps in the key-frames with the help of extracted depth maps.

Though the proposed method can segment the prominent, convex, and elevated polyps that need immediate medical attention, the other polyp structures can not be neglected. They may progress to the malignant stage if not diagnosed early. Therefore, our subsequent studies will focus on segmenting any polyp structures. The following method encapsulates polyps' contextual and spatial information for polyp vs. non-polyp discrimination.

The second segmentation approach can produce better segmentation performance with a reduced computational load than the pixel-based method. The texture and color information is embedded into an MRF framework to extract the polyp pixels' global and local information. The MRF parameter $\boldsymbol{\beta}$ has been made adaptive based on the dominant image characteristics in our approach. The proposed endoscopic polyp segmentation method is unsupervised, whereas most state-of-the-art techniques are supervised. Our algorithm sometimes misses partially occluded polyps and polyps that are texturally not discriminable from the complex colonoscopy frame background. Similarly, polyps that are very small in size may result in an over-segmentation outcome by our method.

In the subsequent method, shape compactness prior is used to detect the ROI of arbitrary polyp shapes. Salient object detection is a method to accentuate the object of interest. A deep U-Net-based CNN model is used for saliency detection in our approach. Segmentation is formulated by energy terms given by a probability map and compactness prior. To solve the energy function, ADMM is used, which is efficient and fast. A series of comparative tests using the publicly available dataset of colonoscopy polyps are used to assess the effectiveness of our method. Compared to the state-of-the-art methods, the experimental results suggest that the proposed technique has competitive performance. Our segmentation method can be used for colonoscopy image analysis. In the future, other priors can also be considered for our model to detect salient regions and may help in better polyp segmentation.

(iii) **Polyp classification:**

Developing a robust and modality-independent polyp classifier is vital for a reliable diagnosis. The colonoscopy images are often susceptible to photometric and geometric transformations. Also, the subtle difference in features makes extracting the discriminating features difficult. A highly efficient and generalizable classifier is indispensable in polyp analysis. Also, the classifier should be data-efficient. Developing supervised models using large annotated images from the classes is a constraint in colonoscopy. In such scenarios, devising a highly effective polyp classifier is challenging. This thesis proposes three methods to address all the problems to some degree.

Our first approach proposes using the shape and texture features of the polyps to classify the stages of dysplasia. The local polyp shape features are extracted using PHOG, and the local

texture features are extracted using FWLBP. The endoscopic video frames are prone to affine transformations. So, the characteristics of the polyp may be perceived differently for different ambient conditions. SPM and FD were incorporated with this framework to deal with this problem. A feature ranking algorithm based on fuzzy entropy is adopted for feature selection. The final assessment using different performance matrices establishes the efficacy of the proposed work. This method is also applied to our own dataset, which contains images from all the three modalities, viz., NBI, WL, and Dye. The consistency in performance demonstrates the robustness of our method. Thus, the selected features can serve the purpose of accurate polyp classification for different modalities. Deep learning-based polyp classification methods have a big challenge due to the lack of extensive and publicly available annotated databases. On the contrary, our method can be deployed to both offline and real-time colonoscopic polyp classification. Though the proposed method achieves an acceptable classification accuracy, the learned texture features may not be sufficient to discriminate between polyp classes. The idea is to learn distributed embeddings representation of data points so that contextually similar data points are projected in the nearby region in the low dimensional vector space. In contrast, different data points are projected far away from each other. This technique, therefore, extract such features from the polyp classes that ensure inter-class separability. We propose a triplet network based on siamese architecture, followed by SVM, to achieve this. Additionally, local polyp features using PHOG are extracted and fused with deep features, resulting in improved classification results. The effectiveness of our strategy in a limited data environment is demonstrated by its classification performance on a relatively small dataset. We hope to improve polyp detection and localization in the future by training the network with features that best characterize the polyp clinical manifestations. Further, grading dysplasia in polyps could also allow practitioners to better comprehend pathological situations. The performance of the suggested algorithm outperforms state-of-the-art approaches by a significant margin.

5.2 Contributions

The major contributions of the research work reported in this thesis includes,

(i) Method for the detection of polyps using dominant and discriminating polyp features.

(ii) A tracking-based algorithm for detecting and segmentation of polyps under different back-

grounds.

(iii) Attention based deep feature extraction for real-time polyp detection.

(iv) Methods for the segmentation of polyps by emphasizing global, local, and spatial information.

(v) Polyp detection and classification in an integrated framework.

(vi) Polyp classification approaches using polyp texture, color, shape features—introduction of deep embedding feature and hand-crafted features for improved classification efficiency.

(vii) Developing a robust, efficient, and generalizable classifier in a limited data environment.

(viii) Dysplasia grading in polyps by learning generic and discriminating features from polyp histopathological images in a GAN-based semisupervised framework.

5.3 Directions for Future Work

Based on the outcome of this thesis work, this section provides some of the possible future directions for research.

(i) To provide robustness, the dominant polyp features are used for the detection of polyps. The performance of detection algorithms may be improved by combining the most discriminating polyp features. A better saliency map can achieve a better localization performance.

(ii) To develop a framework that can be used to select the most significant key-frames from the colonoscopy videos for enhanced diagnosis and clinical management.

(iii) For the detection of polyps, we have used attention blocks in a deep learning framework. The detection accuracy may be improved by providing a domain adaptive attention model along with the proposed channel and spatial attention blocks.

(iv) To develop a model that can detect serrated and diminutive polyps in complex backgrounds.

(v) For the segmentation of polyps, we used the texture, color, shape, and AC. Since boundary information is important for perfect delineation, the extraction of the polyp boundary information may be useful. In this context, the boundary-aware feature extraction can be useful, which will extract the local features separated by the learned boundaries.

(vi) Combination of discriminating and generative features from the polyp classes may better characterize the polyps.

(vii) The classification performances are relatively more affected under degraded conditions. The effect of degradation on polyp regions may be reduced by applying a suitable image enhancement method.

(viii) The learning of features in limited data environment is crucial. Therefore, learning from labeled data and unlabeled data would be helpful in generalization of the system.

List of Publications

Journal Publications

1. P. Sasmal, M. K. Bhuyan, S. Gupta and Y. Iwahori, "Detection of Polyps in Colonoscopic Videos Using Saliency Map-Based Modified Particle Filter," **IEEE Transactions on Instrumentation and Measurement**, vol. 70, pp. 1-9, 2021, Art no. 5011209, doi: 10.1109/TIM.2021.3082315.

2. P. Sasmal, M. K. Bhuyan, Y. Iwahori and K. Kasugai, "Colonoscopic Polyp Classification Using Local Shape and Texture Features," **IEEE Access**, vol. 9, pp. 92629-92639, 2021, doi: 10.1109/ACCESS.2021.3092263.

3. Pradipta Sasmal, M.K. Bhuyan, Soumayan Dutta, and Yuji Iwahori, "An Unsupervised Approach of Colonic Polyp Segmentation using Adaptive Markov Random Fields," **Pattern Recognition Letters (Elsevier)**, vol. 154, pp. 7-15, 2022.

4. P. Sasmal, A. Paul, M. K. Bhuyan, Y. Iwahori and K. Kasugai, "Extraction of Key-Frames From Endoscopic Videos by Using Depth Information," **IEEE Access**, vol. 9, pp. 153004-153011, 2021, doi: 10.1109/ACCESS.2021.3126835.

5. Pradipta Sasmal, M.K. Bhuyan, and Yuji Iwahori,"A saliency map-guided shape compactness for segmentation of polyps in colonoscopy images," **Signal, Image and Video Processing (Springer)**, page: 1-7, SIViP (2022), https://doi.org/10.1007/s11760-022-02195-2.

6. Pradipta Sasmal, M.K. Bhuyan, and Yuji Iwahori, "An Attention YOLO Framework for Automated Detection and Classification of Polyps in Colonoscopy Videos," **IEEE Transactions on Instrumentation and Measurement (IEEE), (Submitted)**.

7. Pradipta Sasmal, Vanshali Sharma, M.K. Bhuyan, and Yuji Iwahori, "A semisupervised GAN for polyp classification using Histopathological Images," **IEEE/ACM Transactions on Computational Biology and Bioinformatics, (Under Review)**.

8. Vanshali Sharma, Pradipta Sasmal, M.K. Bhuyan, PK Das, and Kunio Kasugai "A Multi-scale Attention Framework for Polyp Localization and Keyframe Extraction from Colonoscopy Videos", **IEEE Transactions on Medical Imaging, (To be Submitted)**.

Conference Publications

1. Pradipta Sasmal, M. K. Bhuyan, Sourav Sonowal, Yuji Iwahori, and Kunio Kasugai. "Improved Endoscopic Polyp Classification using GAN Generated Synthetic Data Augmentation," *in IEEE Applied Signal Processing Conference (ASPCON),* pp. 247-251. IEEE, 2020.

2. Pradipta Sasmal, M. K. Bhuyan, Kangkana Bora, Yuji Iwahori, and Kunio Kasugai. "Colonoscopic Image Polyp Classification Using Texture Features," *in International Conference on Pattern Recognition and Machine Intelligence,* pp. 96-101. Springer, Cham, 2019.

3. Pradipta Sasmal, Yuji Iwahori, M. K. Bhuyan, and Kunio Kasugai. "Classification of Polyps in Capsule Endoscopic Images using CNN," *in IEEE Applied Signal Processing Conference (ASPCON),* pp. 253-256. IEEE, 2018.

4. Pradipta Sasmal, Yuji Iwahori, M. K. Bhuyan, and Kunio Kasugai. "Active contour segmentation of polyps in capsule endoscopic images," *in IEEE International Conference on Signals and Systems (ICSigSys),* pp. 201-204, 2018.

5. Soumayan Dutta, Pradipta Sasmal, M. K. Bhuyan, and Yuji Iwahori. "Automatic Segmentation of Polyps in Endoscopic Image Using Level-Set Formulation," *in IEEE International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET),* pp. 1-5., 2018.

6. Ankur Deka, Yuji Iwahori, M. K. Bhuyan, Pradipta Sasmal, and Kunio Kasugai. "Dense 3D Reconstruction of Endoscopic Polyp," *in International Conference on Bioimaging (BIOIMAGING),* pp. 159-166. 2018.

7. Shashwata Gupta, M. K. Bhuyan, and Pradipta Sasmal "Occlusion Robust Object Tracking with Modified Particle Filter Framework," *in IEEE Applied Signal Processing Conference (ASPCON),* pp. 257-261., 2020.

8. Vanshali Sharma, Pradipta Sasmal, M.K. Bhuyan, and PK Das, "Key-frame selection from Colonoscopy videos for Enhanced Polyp Detection," *International Conference on Pattern Recognition (ICPR) 2022.* (Submitted)

Bibliography

[1] M. Arnold, M. S. Sierra, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global patterns and trends in colorectal cancer incidence and mortality," *Gut*, vol. 66, no. 4, pp. 683–691, 2017.

[2] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[3] A. Wiegering, S. Ackermann, J. Riegel, U. A. Dietz, O. Götze, C.-T. Germer, and I. Klein, "Improved survival of patients with colon cancer detected by screening colonoscopy," *International journal of colorectal disease*, vol. 31, no. 5, pp. 1039–1045, 2016.

[4] H. Messmann, *Atlas of Colonoscopy: Techniques-Diagnosis-Interventional Procedures*. Thieme, 2006.

[5] K. Bibbins-Domingo, D. C. Grossman, S. J. Curry, K. W. Davidson, J. W. Epling, F. A. García, M. W. Gillman, D. M. Harper, A. R. Kemper, A. H. Krist *et al.*, "Screening for colorectal cancer: Us preventive services task force recommendation statement," *Jama*, vol. 315, no. 23, pp. 2564–2575, 2016.

[6] A. J. Markowitz and S. J. Winawer, "Management of colorectal polyps," *CA: a cancer journal for clinicians*, vol. 47, no. 2, pp. 93–112, 1997.

[7] J. L. Vleugels, Y. Hazewinkel, P. Fockens, and E. Dekker, "Natural history of diminutive and small colorectal polyps: a systematic literature review," *Gastrointestinal endoscopy*, vol. 85, no. 6, pp. 1169–1176, 2017.

[8] J. L. Vleugels, Y. Hazewinkel, and E. Dekker, "Morphological classifications of gastrointestinal lesions," *Best Practice & Research Clinical Gastroenterology*, vol. 31, no. 4, pp. 359–367, 2017.

[9] A. M. Buchner, "The role of chromoendoscopy in evaluating colorectal dysplasia," *Gastroenterology & hepatology*, vol. 13, no. 6, p. 336, 2017.

[10] H.-G. Lee, M.-K. Choi, B.-S. Shin, and S.-C. Lee, "Reducing redundancy in wireless capsule endoscopy videos," *Computers in biology and medicine*, vol. 43, no. 6, pp. 670–682, 2013.

[11] C. J. Kahi, D. G. Hewett, D. L. Norton, G. J. Eckert, and D. K. Rex, "Prevalence and variable detection of proximal colon serrated polyps during screening colonoscopy," *Clinical Gastroenterology and Hepatology*, vol. 9, no. 1, pp. 42–46, 2011.

[12] D. K. Rex, "Narrow-band imaging without optical magnification for histologic analysis of colorectal polyps," *Gastroenterology*, vol. 136, no. 4, pp. 1174–1181, 2009.

[13] Y. Wada, S.-e. Kudo, H. Kashida, N. Ikehara, H. Inoue, F. Yamamura, K. Ohtsuka, and S. Hamatani, "Diagnosis of colorectal lesions with the magnifying narrow-band imaging system," *Gastrointestinal endoscopy*, vol. 70, no. 3, pp. 522–531, 2009.

[14] C.-G. Guo, R. Ji, and Y.-Q. Li, "Accuracy of i-scan for optical diagnosis of colonic polyps: a meta-analysis," *PloS one*, vol. 10, no. 5, p. e0126237, 2015.

[15] J. Pohl, M. Nguyen-Tat, O. Pech, A. May, T. Rabenstein, and C. Ell, "Computed virtual chromoendoscopy for classification of small colorectal lesions: a prospective comparative study," *American Journal of Gastroenterology*, vol. 103, no. 3, pp. 562–569, 2008.

[16] M. Min, S. Su, W. He, Y. Bi, Z. Ma, and Y. Liu, "Computer-aided diagnosis of colorectal polyps using linked color imaging colonoscopy to predict histology," *Scientific reports*, vol. 9, no. 1, pp. 1–8, 2019.

[17] A. Bond and S. Sarkar, "New technologies and techniques to improve adenoma detection in colonoscopy," *World journal of gastrointestinal endoscopy*, vol. 7, no. 10, p. 969, 2015.

[18] T. Kuiper, W. A. Marsman, J. M. Jansen, E. J. van Soest, Y. C. Haan, G. J. Bakker, P. Fockens, and E. Dekker, "Accuracy for optical diagnosis of small colorectal polyps in nonacademic settings," *Clinical Gastroenterology and Hepatology*, vol. 10, no. 9, pp. 1016–1020, 2012.

[19] D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. De Lange, P. Halvorsen, and H. D. Johansen, "Resunet++: An advanced architecture for medical image segmentation," in *2019 IEEE International Symposium on Multimedia (ISM)*. IEEE, 2019, pp. 225–2255.

[20] M. L. Giger, K. Doi, and H. MacMahon, "Image feature analysis and computer-aided diagnosis in digital radiography. 3. automated detection of nodules in peripheral lung fields," *Medical Physics*, vol. 15, no. 2, pp. 158–166, 1988.

[21] M. Alagappan, J. R. G. Brown, Y. Mori, and T. M. Berzin, "Artificial intelligence in gastrointestinal endoscopy: The future is almost here," *World journal of gastrointestinal endoscopy*, vol. 10, no. 10, p. 239, 2018.

[22] Y. Mori, S.-e. Kudo, T. M. Berzin, M. Misawa, and K. Takeda, "Computer-aided diagnosis for colonoscopy," *Endoscopy*, vol. 49, no. 08, pp. 813–819, 2017.

[23] M. Liedlgruber and A. Uhl, "Computer-aided decision support systems for endoscopy in the gastrointestinal tract: a review," *IEEE reviews in biomedical engineering*, vol. 4, pp. 73–88, 2011.

[24] R. Djinbachian, A.-J. Dubé, and D. von Renteln, "Optical diagnosis of colorectal polyps: recent developments," *Current treatment options in gastroenterology*, vol. 17, no. 1, pp. 99–114, 2019.

[25] S.-e. Kudo, Y. Mori, M. Misawa, K. Takeda, T. Kudo, H. Itoh, M. Oda, and K. Mori, "Artificial intelligence and colonoscopy: Current status and future perspectives," *Digestive Endoscopy*, vol. 31, no. 4, pp. 363–371, 2019.

[26] C. A. Barbano, D. Perlo, E. Tartaglione, A. Fiandrotti, L. Bertero, P. Cassoni, and M. Grangetto, "Unitopatho, a labeled histopathological dataset for colorectal polyps classification and adenoma dysplasia grading," *arXiv preprint arXiv:2101.09991*, 2021.

[27] S. B. Gokturk, C. Tomasi, B. Acar, C. F. Beaulieu, D. S. Paik, R. J. Jeffrey, J. Yee, and S. Napel, "A statistical 3-d pattern processing method for computer-aided detection of polyps in ct colonography," *IEEE Transactions on Medical Imaging*, vol. 20, no. 12, pp. 1251–1260, 2001.

[28] D. S. Paik, C. Beaulieu, R. Jeffrey, C. Karadi, and S. Napel, "Detection of polyps in ct colonography: A comparison of a computer-aided detection algorithm to 3d visualization methods," in *Radiology*, vol. 213. RADIOLOGICAL SOC NORTH AMER 20TH AND NORTHAMPTON STS, EASTON, PA 18042 USA, 1999, pp. 197–197.

[29] H. Yoshida, Y. Masutani, P. MacEneaney, K. Doi, Y. Kim, and A. Dachman, "Detection of colonic polyps in ct colonography based on geometric features," in *Radiology*, vol. 217. RADIOLOGICAL SOC NORTH AMER 20TH AND NORTHAMPTON STS, EASTON, PA 18042 USA, 2000, pp. 582–582.

[30] J. Näppi and H. Yoshida, "Automated detection of polyps with ct colonography: evaluation of volumetric features for reduction of false-positive findings," *Academic Radiology*, vol. 9, no. 4, pp. 386–397, 2002.

[31] S. A. Karkanis, D. K. Iakovidis, D. E. Maroulis, D. A. Karras, and M. Tzivras, "Computer-aided tumor detection in endoscopic video using color wavelet features," *IEEE transactions on information technology in biomedicine*, vol. 7, no. 3, pp. 141–152, 2003.

[32] V. Kodogiannis and M. Boulougoura, "An adaptive neurofuzzy approach for the diagnosis in wireless capsule endoscopy imaging," *International Journal of Information Technology*, vol. 13, no. 1, pp. 46–56, 2007.

[33] B. Li, M. Q.-H. Meng, and L. Xu, "A comparative study of shape features for polyp detection in wireless capsule endoscopy images," in *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2009, pp. 3731–3734.

[34] B. Li, Y. Fan, M. Q.-H. Meng, and L. Qi, "Intestinal polyp recognition in capsule endoscopy images using color and shape features," in *2009 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE, 2009, pp. 1490–1494.

[35] A. Karargyris and N. Bourbakis, "Identification of polyps in wireless capsule endoscopy videos using log gabor filters," in *2009 IEEE/NIH Life Science Systems and Applications Workshop*. IEEE, 2009, pp. 143–147.

[36] S. Hwang and M. E. Celebi, "Polyp detection in wireless capsule endoscopy videos based on image segmentation and geometric feature," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 678–681.

[37] R. D. Nawarathna, J. Oh, X. Yuan, J. Lee, and S. J. Tang, "Abnormal image detection using texton method in wireless capsule endoscopy videos," in *International Conference on Medical Biometrics*. Springer, 2010, pp. 153–162.

[38] R. Nawarathna, J. Oh, J. Muthukudage, W. Tavanapong, J. Wong, P. C. De Groen, and S. J. Tang, "Abnormal image detection in endoscopy videos using a filter bank and local binary patterns," *Neurocomputing*, vol. 144, pp. 70–91, 2014.

[39] P. N. Figueiredo, I. N. Figueiredo, S. Prasath, and R. Tsai, "Automatic polyp detection in pillcam colon 2 capsule images and videos: Preliminary feasibility report," *Diagnostic and Therapeutic Endoscopy*, vol. 2011, 2011.

[40] F. Condessa and J. Bioucas-Dias, "Segmentation and detection of colorectal polyps using local polynomial approximation," in *International Conference Image Analysis and Recognition*. Springer, 2012, pp. 188–197.

[41] E. David, R. Boia, A. Malaescu, and M. Carnu, "Automatic colon polyp detection in endoscopic capsule images," in *International Symposium on Signals, Circuits and Systems ISSCS2013*. IEEE, 2013, pp. 1–4.

[42] I. N. Figueiredo, S. Kumar, and P. N. Figueiredo, "An intelligent system for polyp detection in wireless capsule endoscopy images," *Computational Vision and Medical Image Processing IV: VIPIMAGE*, vol. 2013, pp. 229–235, 2013.

[43] Q. Zhao, T. Dassopoulos, G. Mullin, G. Hager, M. Q. Meng, and R. Kumar, "Towards integrating temporal information in capsule endoscopy image analysis," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2011, pp. 6627–6630.

[44] B. Li and M. Q.-H. Meng, "Automatic polyp detection for wireless capsule endoscopy images," *Expert Systems with Applications*, vol. 39, no. 12, pp. 10 952–10 958, 2012.

[45] Y. Yuan and M. Q.-H. Meng, "Polyp classification based on bag of features and saliency in wireless capsule endoscopy," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2014, pp. 3930–3935.

[46] Y. Yuan, B. Li, and M. Q.-H. Meng, "Improved bag of feature for automatic polyp detection in wireless capsule endoscopy images," *IEEE Transactions on automation science and engineering*, vol. 13, no. 2, pp. 529–535, 2015.

[47] Y. Wang, W. Tavanapong, J. Wong, J. H. Oh, and P. C. De Groen, "Polyp-alert: Near real-time feedback during colonoscopy," *Computer methods and programs in biomedicine*, vol. 120, no. 3, pp. 164–179, 2015.

[48] S. Krishnan, X. Yang, K. Chan, S. Kumar, and P. Goh, "Intestinal abnormality detection from endoscopic images," in *Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Vol. 20 Biomedical Engineering Towards the Year 2000 and Beyond (Cat. No. 98CH36286)*, vol. 2. IEEE, 1998, pp. 895–898.

[49] P. Wang, S. M. Krishnan, C. Kugean, and M. Tjoa, "Classification of endoscopic images based on texture and neural network," in *2001 Conference Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 4. IEEE, 2001, pp. 3691–3695.

[50] G. D. Magoulas, V. P. Plagianakos, and M. N. Vrahatis, "Improved neural networkbased interpretation of colonoscopy images through on-line learning and evolution," *European Network of Excellence on Intelligent Technologies for Smart Adaptive Systems*, pp. 38–43, 2001.

[51] J. Kang and R. Doraiswami, "Real-time image processing system for endoscopic applications," in *CCECE 2003-Canadian Conference on Electrical and Computer Engineering. Toward a Caring and Humane Technology (Cat. No. 03CH37436)*, vol. 3. IEEE, 2003, pp. 1469–1472.

[52] M. P. Tjoa and S. M. Krishnan, "Feature extraction for the analysis of colon status from the endoscopic images," *BioMedical Engineering OnLine*, vol. 2, no. 1, pp. 1–17, 2003.

[53] G. Magoulas, V. Plagianakos, D. Tasoulis, and M. Vrahatis, "Tumor detection in colonoscopy using the unsupervised k-windows clustering algorithm and neural networks," in *Fourth European Symposium on "Biomedical Engineering*, 2004.

[54] B. V. Dhandra, R. Hegadi, M. Hangarge, and V. S. Malemath, "Analysis of abnormality in endoscopic images using combined hsi color space and watershed segmentation," in *18th International Conference on Pattern Recognition (ICPR'06)*, vol. 4. IEEE, 2006, pp. 695–698.

[55] D. K. Iakovidis, D. E. Maroulis, and S. A. Karkanis, "An intelligent system for automatic detection of gastrointestinal adenomas in video endoscopy," *Computers in biology and medicine*, vol. 36, no. 10, pp. 1084–1103, 2006.

[56] S. Hwang, J. Oh, W. Tavanapong, J. Wong, and P. C. De Groen, "Polyp detection in colonoscopy video using elliptical shape feature," in *2007 IEEE International Conference on Image Processing*, vol. 2. IEEE, 2007, pp. II–465.

[57] L. A. Alexandre, J. Casteleiro, and N. Nobreinst, "Polyp detection in endoscopic video using svms," in *European Conference on Principles of Data Mining and Knowledge Discovery*. Springer, 2007, pp. 358–365.

[58] D.-C. Cheng, W.-C. Ting, Y.-F. Chen, Q. Pu, and X. Jiang, "Colorectal polyps detection using texture features and support vector machine," in *International Conference on Mass Data Analysis of Images and Signals in Medicine, Biotechnology, and Chemistry*. Springer, 2008, pp. 62–72.

[59] S. Ameling, S. Wirth, D. Paulus, G. Lacey, and F. Vilarino, "Texture-based polyp detection in colonoscopy," in *Bildverarbeitung für die Medizin 2009*. Springer, 2009, pp. 346–350.

[60] S. Y. Park, D. Sargent, I. Spofford, K. G. Vosburgh, A. Yousif *et al.*, "A colon video analysis framework for polyp detection," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 5, pp. 1408–1418, 2012.

[61] J. Bernal, J. Sánchez, and F. Vilarino, "Towards automatic polyp detection with a polyp appearance model," *Pattern Recognition*, vol. 45, no. 9, pp. 3166–3182, 2012.

[62] Y. Wang, W. Tavanapong, J. Wong, J. Oh, and P. C. De Groen, "Part-based multiderivative edge cross-sectional profiles for polyp detection in colonoscopy," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 4, pp. 1379–1389, 2013.

[63] N. Tajbakhsh, C. Chi, S. R. Gurudu, and J. Liang, "Automatic polyp detection from learned boundaries," in *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2014, pp. 97–100.

[64] Q. Zhao and M. Q.-H. Meng, "Polyp detection in wireless capsule endoscopy images using novel color texture features," in *2011 9th World Congress on Intelligent Control and Automation*. IEEE, 2011, pp. 948–952.

[65] Q. Zhao, T. Dassopoulos, G. E. Mullin, M. Q.-H. Meng, and R. Kumar, "A decision fusion strategy for polyp detection in capsule endoscopy," in *Medicine Meets Virtual Reality 19*. IOS Press, 2012, pp. 559–565.

[66] S. Hwang, "Bag-of-visual-words approach to abnormal image detection in wireless capsule endoscopy videos," in *International Symposium on Visual Computing*. Springer, 2011, pp. 320–327.

[67] Y. Yuan and M. Q.-H. Meng, "A novel feature for polyp detection in wireless capsule endoscopy images," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2014, pp. 5010–5015.

[68] J. Jia, S. Sun, T. Chen, and P. Wang, "Accurate and efficient polyp detection in wireless capsule endoscopy images," Aug. 29 2017, uS Patent 9,743,824.

[69] M. Zhou, G. Bao, Y. Geng, B. Alkandari, and X. Li, "Polyp detection and radius measurement in small intestine using video capsule endoscopy," in *2014 7th International Conference on Biomedical Engineering and Informatics*. IEEE, 2014, pp. 237–241.

[70] L. Gueye, S. Yildirim-Yayilgan, F. A. Cheikh, and I. Balasingham, "Automatic detection of colonoscopic anomalies using capsule endoscopy," in *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2015, pp. 1061–1064.

[71] D. Jha, S. Ali, N. K. Tomar, H. D. Johansen, D. Johansen, J. Rittscher, M. A. Riegler, and P. Halvorsen, "Real-time polyp detection, localization and segmentation in colonoscopy using deep learning," *Ieee Access*, vol. 9, pp. 40 496–40 510, 2021.

[72] P. Wang, X. Xiao, J. R. G. Brown, T. M. Berzin, M. Tu, F. Xiong, X. Hu, P. Liu, Y. Song, D. Zhang *et al.*, "Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy," *Nature biomedical engineering*, vol. 2, no. 10, pp. 741–748, 2018.

[73] G. Urban, P. Tripathi, T. Alkayali, M. Mittal, F. Jalali, W. Karnes, and P. Baldi, "Deep learning localizes and identifies polyps in real time with 96% accuracy in screening colonoscopy," *Gastroenterology*, vol. 155, no. 4, pp. 1069–1078, 2018.

[74] J. Y. Lee, J. Jeong, E. M. Song, C. Ha, H. J. Lee, J. E. Koo, D.-H. Yang, N. Kim, and J.-S. Byeon, "Real-time detection of colon polyps during colonoscopy using deep learning: systematic validation with four independent datasets," *Scientific reports*, vol. 10, no. 1, pp. 1–9, 2020.

[75] O. F. Ahmad, A. S. Soares, E. Mazomenos, P. Brandao, R. Vega, E. Seward, D. Stoyanov, M. Chand, and L. B. Lovat, "Artificial intelligence and computer-aided diagnosis in colonoscopy: current evidence and future directions," *The lancet Gastroenterology & hepatology*, vol. 4, no. 1, pp. 71–80, 2019.

[76] Q. Li, G. Yang, Z. Chen, B. Huang, L. Chen, D. Xu, X. Zhou, S. Zhong, H. Zhang, and T. Wang, "Colorectal polyp segmentation using a fully convolutional neural network," in *2017 10th international congress on image and signal processing, biomedical engineering and informatics (CISP-BMEI)*. IEEE, 2017, pp. 1–5.

[77] D. Vázquez, J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, A. M. López, A. Romero, M. Drozdzal, and A. Courville, "A benchmark for endoluminal scene segmentation of colonoscopy images," *Journal of healthcare engineering*, vol. 2017, 2017.

[78] J. Bernal, N. Tajkbaksh, F. J. Sanchez, B. J. Matuszewski, H. Chen, L. Yu, Q. Angermann, O. Romain, B. Rustad, I. Balasingham *et al.*, "Comparative validation of polyp detection methods in video colonoscopy: results from the miccai 2015 endoscopic vision challenge," *IEEE transactions on medical imaging*, vol. 36, no. 6, pp. 1231–1249, 2017.

[79] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[80] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?" *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1299–1312, 2016.

[81] Y. Shin, H. A. Qadir, L. Aabakken, J. Bergsland, and I. Balasingham, "Automatic colon polyp detection using region based deep cnn and post learning approaches," *IEEE Access*, vol. 6, pp. 40 950–40 962, 2018.

[82] Y. Shin, H. A. Qadir, and I. Balasingham, "Abnormal colon polyp image synthesis using conditional adversarial networks for improved detection performance," *IEEE Access*, vol. 6, pp. 56 007–56 017, 2018.

[83] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[84] M. Yamada, Y. Saito, H. Imaoka, M. Saiko, S. Yamada, H. Kondo, H. Takamaru, T. Sakamoto, J. Sese, A. Kuchiba *et al.*, "Development of a real-time endoscopic image diagnosis support system using deep learning technology in colonoscopy," *Scientific reports*, vol. 9, no. 1, pp. 1–9, 2019.

[85] B. Taha, J. Dias, and N. Werghi, "Convolutional neural networkasa feature extractor for automatic polyp detection," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 2060–2064.

[86] Y. Shin and I. Balasingham, "Comparison of hand-craft feature based svm and cnn based deep learning framework for automatic polyp classification," in *2017 39th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE, 2017, pp. 3277–3280.

[87] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "On the necessity of fine-tuned convolutional neural networks for medical imaging," in *Deep Learning and Convolutional Neural Networks for Medical Image Computing*. Springer, 2017, pp. 181–193.

[88] Z. Yuan, M. IzadyYazdanabadi, D. Mokkapati, R. Panvalkar, J. Y. Shin, N. Tajbakhsh, S. Gurudu, and J. Liang, "Automatic polyp detection in colonoscopy videos," in *Medical Imaging 2017: Image Processing*, vol. 10133. International Society for Optics and Photonics, 2017, p. 101332K.

[89] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automatic polyp detection in colonoscopy videos using an ensemble of convolutional neural networks," in *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2015, pp. 79–83.

[90] ——, "A comprehensive computer-aided polyp detection system for colonoscopy videos," in *International Conference on Information Processing in Medical Imaging*. Springer, 2015, pp. 327–338.

[91] S. Axyonov, J. Liang, K. Kostin, A. V. Zamyatin *et al.*, "Advanced pattern recognition and deep learning for colon polyp detection," in *Distributed computer and communication networks: control, computation, communications (DCCN-2016)*, 2016, pp. 27–34.

[92] M. Akbari, M. Mohrekesh, S. Rafiei, S. R. Soroushmehr, N. Karimi, S. Samavi, and K. Najarian, "Classification of informative frames in colonoscopy videos using convolutional neural networks with binarized weights," in *2018 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE, 2018, pp. 65–68.

[93] H. Itoh, H. R. Roth, L. Lu, M. Oda, M. Misawa, Y. Mori, S.-e. Kudo, and K. Mori, "Towards automated colonoscopy diagnosis: binary polyp size estimation via unsupervised depth learning," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2018, pp. 611–619.

[94] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[95] M. Misawa, S.-e. Kudo, Y. Mori, T. Cho, S. Kataoka, A. Yamauchi, Y. Ogawa, Y. Maeda, K. Takeda, K. Ichimasa *et al.*, "Artificial intelligence-assisted polyp detection for colonoscopy: initial experience," *Gastroenterology*, vol. 154, no. 8, pp. 2027–2029, 2018.

[96] V. N. Murthy, V. Singh, S. Sun, S. Bhattacharya, T. Chen, and D. Comaniciu, "Cascaded deep decision networks for classification of endoscopic images," in *Medical Imaging 2017: Image Processing*, vol. 10133. International Society for Optics and Photonics, 2017, p. 101332B.

[97] X. Mo, K. Tao, Q. Wang, and G. Wang, "An efficient approach for polyps detection in endoscopic videos based on faster r-cnn," in *2018 24th international conference on pattern recognition (ICPR)*. IEEE, 2018, pp. 3929–3934.

[98] A. Mohammed, S. Yildirim, I. Farup, M. Pedersen, and Ø. Hovde, "Y-net: A deep convolutional neural network for polyp detection," *arXiv preprint arXiv:1806.01907*, 2018.

[99] L. Yu, H. Chen, Q. Dou, J. Qin, and P. A. Heng, "Integrating online and offline three-dimensional deep learning for automated polyp detection in colonoscopy videos," *IEEE journal of biomedical and health informatics*, vol. 21, no. 1, pp. 65–75, 2016.

[100] R. Zhang, Y. Zheng, C. C. Poon, D. Shen, and J. Y. Lau, "Polyp detection during colonoscopy using a regression-based convolutional neural network with a tracker," *Pattern recognition*, vol. 83, pp. 209–219, 2018.

[101] M. Billah, S. Waheed, and M. M. Rahman, "An automatic gastrointestinal polyp detection system in video endoscopy using fusion of color wavelet and convolutional neural network features," *International journal of biomedical imaging*, vol. 2017, 2017.

[102] Y. Zheng, R. Zhang, R. Yu, Y. Jiang, T. W. Mak, S. H. Wong, J. Y. Lau, and C. C. Poon, "Localisation of colorectal polyps by convolutional neural network features learnt from white light and narrow band endoscopic images of multiple databases," in *2018 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE, 2018, pp. 4142–4145.

[103] K. Pogorelov, M. Riegler, S. L. Eskeland, T. de Lange, D. Johansen, C. Griwodz, P. T. Schmidt, and P. Halvorsen, "Efficient disease detection in gastrointestinal videos–global features versus neural networks," *Multimedia Tools and Applications*, vol. 76, no. 21, pp. 22 493–22 525, 2017.

[104] M. Gadermayr, A. Uhl, and A. Vécsei, "Fully automated decision support systems for celiac disease diagnosis," *IRBM*, vol. 37, no. 1, pp. 31–39, 2016.

[105] I. N. Figueiredo, L. Pinto, P. N. Figueiredo, and R. Tsai, "Unsupervised segmentation of colonic polyps in narrow-band imaging data based on manifold representation of images and wasserstein distance," *Biomedical Signal Processing and Control*, vol. 53, p. 101577, 2019.

[106] W. Cao, J. Zheng, D. Xiang, S. Ding, H. Sun, X. Yang, Z. Liu, and Y. Dai, "Edge and neighborhood guidance network for 2d medical image segmentation," *Biomedical Signal Processing and Control*, vol. 69, p. 102856, 2021.

[107] D.-C. Cheng, W.-C. Ting, Y.-F. Chen, Q. Pu, and X. Jiang, "Colorectal polyps detection using texture features and support vector machine," in *International Conference on Mass Data Analysis of Images and Signals in Medicine, Biotechnology, and Chemistry*. Springer, 2008, pp. 62–72.

[108] D. K. Iakovidis, D. E. Maroulis, S. A. Karkanis, and A. Brokos, "A comparative study of texture features for the discrimination of gastric polyps in endoscopic video," in *Computer-Based Medical Systems, 2005. Proceedings. 18th IEEE Symposium on*. IEEE, 2005, pp. 575–580.

[109] P. Sasmal, Y. Iwahori, M. Bhuyan, and K. Kasugai, "Active contour segmentation of polyps in capsule endoscopic images," in *2018 International Conference on Signals and Systems (ICSigSys)*. IEEE, 2018, pp. 201–204.

[110] M. Ganz, X. Yang, and G. Slabaugh, "Automatic segmentation of polyps in colonoscopic narrow-band imaging data," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 8, pp. 2144–2151, 2012.

[111] M. Prastawa, E. Bullitt, S. Ho, and G. Gerig, "A brain tumor segmentation framework based on outlier detection," *Medical image analysis*, vol. 8, no. 3, pp. 275–283, 2004.

[112] P. Brandao, E. Mazomenos, G. Ciuti, R. Caliò, F. Bianchi, A. Menciassi, P. Dario, A. Koulaouzidis, A. Arezzo, and D. Stoyanov, "Fully convolutional neural networks for polyp segmentation in colonoscopy," in *Medical Imaging 2017: Computer-Aided Diagnosis*, vol. 10134. International Society for Optics and Photonics, 2017, p. 101340F.

[113] Q. Nguyen and S.-W. Lee, "Colorectal segmentation using multiple encoder-decoder network in colonoscopy images," in *2018 IEEE first international conference on artificial intelligence and knowledge engineering (AIKE)*. IEEE, 2018, pp. 208–211.

[114] Y. B. Guo and B. Matuszewski, "Giana polyp segmentation with fully convolutional dilation neural networks," in *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*. SCITEPRESS-Science and Technology Publications, 2019, pp. 632–641.

[115] K. Pogorelov, O. Ostroukhova, M. Jeppsson, H. Espeland, C. Griwodz, T. de Lange, D. Johansen, M. Riegler, and P. Halvorsen, "Deep learning and hand-crafted feature based approaches for polyp detection in medical videos," in *2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 2018, pp. 381–386.

[116] D. Banik, K. Roy, D. Bhattacharjee, M. Nasipuri, and O. Krejcar, "Polyp-net: A multimodel fusion network for polyp segmentation," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–12, 2020.

[117] X. Yang, Q. Wei, C. Zhang, K. Zhou, L. Kong, and W. Jiang, "Colon polyp detection and segmentation based on improved mrcnn," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–10, 2020.

[118] S. Gross, M. Kennel, T. Stehle, J. Wulff, J. Tischendorf, C. Trautwein, and T. Aach, "Polyp segmentation in nbi colonoscopy," in *Bildverarbeitung für die Medizin 2009*. Springer, 2009, pp. 252–256.

[119] M. Breier, S. Gross, A. Behrens, T. Stehle, and T. Aach, "Active contours for localizing polyps in colonoscopic nbi image data," in *Medical Imaging 2011: Computer-Aided Diagnosis*, vol. 7963. International Society for Optics and Photonics, 2011, p. 79632M.

[120] K. Wickstrøm, M. Kampffmeyer, and R. Jenssen, "Uncertainty modeling and interpretability in convolutional neural networks for polyp segmentation," in *2018 ieee 28th international workshop on machine learning for signal processing (mlsp)*. IEEE, 2018, pp. 1–6.

[121] L. Zhang, S. Dolwani, and X. Ye, "Automated polyp segmentation in colonoscopy frames using fully convolutional neural network and textons," in *Annual Conference on Medical Image Understanding and Analysis*. Springer, 2017, pp. 707–717.

[122] W.-T. Xiao, L.-J. Chang, and W.-M. Liu, "Semantic segmentation of colorectal polyps with deeplab and lstm networks," in *2018 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*. IEEE, 2018, pp. 1–2.

[123] D. Banik, D. Bhattacharjee, and M. Nasipuri, "A multi-scale patch-based deep learning system for polyp segmentation," in *Advanced Computing and Systems for Security*. Springer, 2020, pp. 109–119.

[124] T. Stehle, R. Auer, S. Gross, A. Behrens, J. Wulff, T. Aach, R. Winograd, C. Trautwein, and J. 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, p. 72602S.

[125] F. J. Condessa, "Detection and classification of human colorectal polyps," *ST, Lisbon, Portugal*, 2011.

[126] J. J. Fu, Y.-W. Yu, H.-M. Lin, J.-W. Chai, and C. C.-C. Chen, "Feature extraction and pattern classification of colorectal polyps in colonoscopic imaging," *Computerized medical imaging and graphics*, vol. 38, no. 4, pp. 267–275, 2014.

[127] M. Häfner, M. Liedlgruber, A. Uhl, A. Vécsei, and F. Wrba, "Color treatment in endoscopic image classification using multi-scale local color vector patterns," *Medical image analysis*, vol. 16, no. 1, pp. 75–86, 2012.

[128] P. Mesejo, D. Pizarro, A. Abergel, O. Rouquette, S. Beorchia, L. Poincloux, and A. Bartoli, "Computer-aided classification of gastrointestinal lesions in regular colonoscopy," *IEEE transactions on medical imaging*, vol. 35, no. 9, pp. 2051–2063, 2016.

[129] G. Wimmer, T. Tamaki, J. J. Tischendorf, M. Häfner, S. Yoshida, S. Tanaka, and A. Uhl, "Directional wavelet based features for colonic polyp classification," *Medical image analysis*, vol. 31, pp. 16–36, 2016.

[130] S. Engelhardt, S. Ameling, S. Wirth, and D. Paulus, "Features for classification of polyps in colonoscopy." in *Bildverarbeitung für die Medizin*, vol. 574, 2010, pp. 350–354.

[131] J. M. Aman, R. M. Summers, and J. Yao, "Characterizing colonic detections in ct colonography using curvature-based feature descriptor and bag-of-words model," in *International MICCAI Workshop on Computational Challenges and Clinical Opportunities in Virtual Colonoscopy and Abdominal Imaging*. Springer, 2010, pp. 15–23.

[132] M. F. Byrne, N. Chapados, F. Soudan, C. Oertel, M. L. Pérez, R. Kelly, N. Iqbal, F. Chandelier, and D. K. Rex, "Real-time differentiation of adenomatous and hyperplastic diminutive colorectal polyps during analysis of unaltered videos of standard colonoscopy using a deep learning model," *Gut*, vol. 68, no. 1, pp. 94–100, 2019.

[133] E. Ribeiro, A. Uhl, and M. Häfner, "Colonic polyp classification with convolutional neural networks," in *2016 IEEE 29th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 2016, pp. 253–258.

[134] E. Ribeiro, A. Uhl, G. Wimmer, and M. Häfner, "Exploring deep learning and transfer learning for colonic polyp classification," *Computational and mathematical methods in medicine*, vol. 2016, 2016.

[135] R. Fonollá, F. van der Sommen, R. M. Schreuder, E. J. Schoon, and P. H. de With, "Multi-modal classification of polyp malignancy using cnn features with balanced class augmentation," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 2019, pp. 74–78.

[136] M. Golhar, T. L. Bobrow, M. P. Khoshknab, S. Jit, S. Ngamruengphong, and N. J. Durr, "Improving colonoscopy lesion classification using semi-supervised deep learning," *IEEE Access*, 2020.

[137] P. Sasmal, M. Bhuyan, S. Sonowal, Y. Iwahori, and K. Kasugai, "Improved endoscopic polyp classification using gan generated synthetic data augmentation," in *2020 IEEE Applied Signal Processing Conference (ASPCON)*. IEEE, 2020, pp. 247–251.

[138] P.-J. Chen, M.-C. Lin, M.-J. Lai, J.-C. Lin, H. H.-S. Lu, and V. S. Tseng, "Accurate classification of diminutive colorectal polyps using computer-aided analysis," *Gastroenterology*, vol. 154, no. 3, pp. 568–575, 2018.

[139] X. Zhang, F. Chen, T. Yu, J. An, Z. Huang, J. Liu, W. Hu, L. Wang, H. Duan, and J. Si, "Real-time gastric polyp detection using convolutional neural networks," *PloS one*, vol. 14, no. 3, p. e0214133, 2019.

[140] Y. J. Kim, J. P. Bae, J.-W. Chung, D. K. Park, K. G. Kim, and Y. J. Kim, "New polyp image classification technique using transfer learning of network-in-network structure in endoscopic images," *Scientific Reports*, vol. 11, no. 1, pp. 1–8, 2021.

[141] B. Korbar, A. M. Olofson, A. P. Miraflor, C. M. Nicka, M. A. Suriawinata, L. Torresani, A. A. Suriawinata, and S. Hassanpour, "Deep learning for classification of colorectal polyps on whole-slide images," *Journal of pathology informatics*, vol. 8, 2017.

[142] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[143] J. W. Wei, A. A. Suriawinata, L. J. Vaickus, B. Ren, X. Liu, M. Lisovsky, N. Tomita, B. Abdollahi, A. S. Kim, D. C. Snover *et al.*, "Evaluation of a deep neural network for automated classification of colorectal polyps on histopathologic slides," *JAMA network open*, vol. 3, no. 4, pp. e203 398–e203 398, 2020.

[144] Z. Song, C. Yu, S. Zou, W. Wang, Y. Huang, X. Ding, J. Liu, L. Shao, J. Yuan, X. Gou *et al.*, "Automatic deep learning-based colorectal adenoma detection system and its similarities with pathologists," *BMJ open*, vol. 10, no. 9, p. e036423, 2020.

[145] S. Gross, C. Trautwein, A. Behrens, R. Winograd, S. Palm, H. H. Lutz, R. Schirin-Sokhan, H. Hecker, T. Aach, and J. J. Tischendorf, "Computer-based classification of small colorectal polyps by using narrow-band imaging with optical magnification," *Gastrointestinal endoscopy*, vol. 74, no. 6, pp. 1354–1359, 2011.

[146] A. Hann, B. M. Walter, N. Mehlhase, and A. Meining, "Virtual reality in gi endoscopy: intuitive zoom for improving diagnostics and training," *Gut*, vol. 68, no. 6, pp. 957–959, 2019.

[147] N. Mahmoud, I. Cirauqui, A. Hostettler, C. Doignon, L. Soler, J. Marescaux, and J. Montiel, "Orbslam-based endoscope tracking and 3d reconstruction," in *International workshop on computer-assisted and robotic endoscopy*. Springer, 2016, pp. 72–83.

[148] X. Liu, A. Sinha, M. Ishii, G. D. Hager, A. Reiter, R. H. Taylor, and M. Unberath, "Dense depth estimation in monocular endoscopy with self-supervised learning methods," *IEEE transactions on medical imaging*, vol. 39, no. 5, pp. 1438–1447, 2019.

[149] B. Li, M. Q.-H. Meng, and Q. Zhao, "Wireless capsule endoscopy video summary," in *2010 IEEE International Conference on Robotics and Biomimetics*. IEEE, 2010, pp. 454–459.

[150] D. K. Iakovidis, S. Tsevas, and A. Polydorou, "Reduction of capsule endoscopy reading times by unsupervised image mining," *Computerized Medical Imaging and Graphics*, vol. 34, no. 6, pp. 471–478, 2010.

[151] S. E. F. De Avila, A. P. B. Lopes, A. da Luz Jr, and A. de Albuquerque Araújo, "Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method," *Pattern Recognition Letters*, vol. 32, no. 1, pp. 56–68, 2011.

[152] X.-s. Hua, L. Lu, H.-j. Zhang, and H. District, "A generic framework of user attention model and its application in video summarization," *IEEE Transaction on multimedia*, vol. 7, no. 5, pp. 907–919, 2005.

[153] N. Ejaz, I. Mehmood, and S. W. Baik, "Mrt letter: Visual attention driven framework for hysteroscopy video abstraction," *Microscopy research and technique*, vol. 76, no. 6, pp. 559–563, 2013.

[154] E. Mendi, C. Bayrak, S. Cecen, and E. Ermisoglu, "Content-based management service for medical videos," *Telemedicine and e-Health*, vol. 19, no. 1, pp. 36–41, 2013.

[155] M. Ma, S. Mei, S. Wan, Z. Wang, and D. Feng, "Video summarization via nonlinear sparse dictionary selection," *IEEE Access*, vol. 7, pp. 11 763–11 774, 2019.

[156] S. Wang, Y. Cong, J. Cao, Y. Yang, Y. Tang, H. Zhao, and H. Yu, "Scalable gastroscopic video summarization via similar-inhibition dictionary selection," *Artificial intelligence in medicine*, vol. 66, pp. 1–13, 2016.

[157] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. D. Johansen, "Kvasir-seg: A segmented polyp dataset," in *International Conference on Multimedia Modeling*. Springer, 2020, pp. 451–462.

[158] M. Misawa, S.-e. Kudo, Y. Mori, K. Hotta, K. Ohtsuka, T. Matsuda, S. Saito, T. Kudo, T. Baba, F. Ishida *et al.*, "Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video)," *Gastrointestinal Endoscopy*, vol. 93, no. 4, pp. 960–967, 2021.

[159] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Computerized Medical Imaging and Graphics*, vol. 43, pp. 99–111, 2015.

[160] Q. Angermann, J. Bernal, C. Sánchez-Montes, M. Hammami, G. Fernández-Esparrach, X. Dray, O. Romain, F. J. Sánchez, and A. 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.

[161] M. Ganz, X. Yang, and G. Slabaugh, "Automatic segmentation of polyps in colonoscopic narrow-band imaging data," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 8, pp. 2144–2151, 2012.

[162] Y. Iwahori, A. Hattori, Y. Adachi, M. K. Bhuyan, R. J. Woodham, and K. Kasugai, "Automatic detection of polyp using hessian filter and hog features," *Procedia computer science*, vol. 60, pp. 730–739, 2015.

[163] D. K. Iakovidis, D. E. Maroulis, S. A. Karkanis, and A. Brokos, "A comparative study of texture features for the discrimination of gastric polyps in endoscopic video," in *18th IEEE Symposium on Computer-Based Medical Systems (CBMS'05)*. IEEE, 2005, pp. 575–580.

[164] S. Gupta, M. Bhuyan, and P. Sasmal, "Occlusion robust object tracking with modified particle filter framework," in *2020 IEEE Applied Signal Processing Conference (ASPCON)*. IEEE, 2020, pp. 257–261.

[165] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 7, pp. 1409–1422, 2011.

[166] X. Cao, X. Jiang, X. Li, and P. Yan, "Correlation-based tracking of multiple targets with hierarchical layered structure," *IEEE transactions on cybernetics*, vol. 48, no. 1, pp. 90–102, 2016.

[167] N. Hussain, A. Khan, S. G. Javed, and M. Hussain, "Particle swarm optimization based object tracking using hog features," in *2013 IEEE 9th International Conference on Emerging Technologies (ICET)*. IEEE, 2013, pp. 1–6.

[168] Y. Lu, Y. Wang, X. Tong, Z. Zhao, H. Jia, and J. Kong, "Face tracking in video sequences using particle filter based on skin color model and facial contour," in *2008 Second International Symposium on Intelligent Information Technology Application*, vol. 1. IEEE, 2008, pp. 457–461.

[169] W. Zheng and S. M. Bhandarkar, "A boosted adaptive particle filter for face detection and tracking," in *2006 International Conference on Image Processing*. IEEE, 2006, pp. 2821–2824.

[170] G. Yu, Z. Hu, H. Lu, and W. Li, "Robust object tracking with occlusion handle," *Neural Computing and Applications*, vol. 20, no. 7, pp. 1027–1034, 2011.

[171] H. Yao and F. Zhu, "Multiple features human face tracking based on particle filter," in *2009 Second International Workshop on Computer Science and Engineering*, vol. 2. IEEE, 2009, pp. 121–125.

[172] L. Zhang, Z. Gu, and H. Li, "Sdsp: A novel saliency detection method by combining simple priors," in *2013 IEEE international conference on image processing*. IEEE, 2013, pp. 171–175.

[173] S. Liu and Y. Peng, "A local region-based chan–vese model for image segmentation," *Pattern Recognition*, vol. 45, no. 7, pp. 2769–2779, 2012.

[174] P. Sasmal, M. K. Bhuyan, K. Bora, Y. Iwahori, and K. Kasugai, "Colonoscopic image polyp classification using texture features," in *International Conference on Pattern Recognition and Machine Intelligence*. Springer, 2019, pp. 96–101.

[175] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[176] J. Bernal, J. M. Núñez, F. J. Sánchez, and F. Vilariño, "Polyp segmentation method in colonoscopy videos by means of msa-dova energy maps calculation," in *Workshop on Clinical Image-Based Procedures*. Springer, 2014, pp. 41–49.

[177] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 5, pp. 898–916, 2010.

[178] I. Pacal, D. Karaboga, A. Basturk, B. Akay, and U. Nalbantoglu, "A comprehensive review of deep learning in colon cancer," *Computers in Biology and Medicine*, vol. 126, p. 104003, 2020.

[179] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.

[180] ——, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[181] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[182] I. Pacal and D. Karaboga, "A robust real-time deep learning based automatic polyp detection system," *Computers in Biology and Medicine*, vol. 134, p. 104519, 2021.

[183] I. Pacal, A. Karaman, D. Karaboga, B. Akay, A. Basturk, U. Nalbantoglu, and S. Coskun, "An efficient real-time colonic polyp detection with yolo algorithms trained by using negative samples and large datasets," *Computers in biology and medicine*, vol. 141, p. 105031, 2022.

[184] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[185] R. Faster, "Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 9199, 2015.

[186] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[187] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[188] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International journal of computer vision*, vol. 111, no. 1, pp. 98–136, 2015.

[189] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 781–10 790.

[190] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[191] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, pp. 91–99, 2015.

[192] J. Wan, B. Chen, and Y. Yu, "Polyp detection from colorectum images by using attentive yolov5," *Diagnostics*, vol. 11, no. 12, p. 2264, 2021.

[193] U. Seitz, T. L. Ang, F. Dy, T. Sookpaisal, J. Sadikin, I. Marki, F. Thonke, S. Seewald, S. Bohnacker, A. De Weerth *et al.*, "Colonic polyps and malignant potential–does size matter?" *Gastrointestinal Endoscopy*, vol. 61, no. 5, p. AB264, 2005.

[194] R. Law, A. Das, D. Gregory, S. Komanduri, R. Muthusamy, A. Rastogi, J. Vargo, M. B. Wallace, G. S. Raju, R. Mounzer *et al.*, "Endoscopic resection is cost-effective compared with laparoscopic resection in the management of complex colon polyps: an economic analysis," *Gastrointestinal endoscopy*, vol. 83, no. 6, pp. 1248–1257, 2016.

[195] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Advances in neural information processing systems*, 2014, pp. 2366–2374.

[196] E. Ricci, W. Ouyang, X. Wang, N. Sebe *et al.*, "Monocular depth estimation using multi-scale continuous crfs as sequential deep networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 6, pp. 1426–1440, 2018.

[197] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2002–2011.

[198] R. Garg, V. K. BG, G. Carneiro, and I. Reid, "Unsupervised cnn for single view depth estimation: Geometry to the rescue," in *European Conference on Computer Vision*. Springer, 2016, pp. 740–756.

[199] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 270–279.

[200] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving," in *Proc. CVPR*, pp. 3354–3361.

[201] A. Saxena, M. Sun, and A. Y. Ng, "Learning 3-d scene structure from a single still image," in *2007 IEEE 11th International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.

[202] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[203] K. Lasinger, R. Ranftl, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *arXiv preprint arXiv:1907.01341*, 2019.

[204] S. Engel, X. Zhang, and B. Wandell, "Colour tuning in human visual cortex measured with functional magnetic resonance imaging," *Nature*, vol. 388, no. 6637, p. 68, 1997.

[205] M.-K. Hu, "Visual pattern recognition by moment invariants," *IRE transactions on information theory*, vol. 8, no. 2, pp. 179–187, 1962.

[206] K. Xian, C. Shen, Z. Cao, H. Lu, Y. Xiao, R. Li, and Z. Luo, "Monocular relative depth perception with web stereo data supervision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 311–320.

[207] Y. Iwahori, T. Shinohara, A. Hattori, R. J. Woodham, S. Fukui, M. K. Bhuyan, and K. Kasugai, "Automatic polyp detection in endoscope images using a hessian filter." in *MVA*, 2013, pp. 21–24.

[208] M. A. Khan, M. A. Khan, F. Ahmed, M. Mittal, L. M. Goyal, D. J. Hemanth, and S. C. Satapathy, "Gastrointestinal diseases segmentation and classification based on duo-deep architectures," *Pattern Recognition Letters*, vol. 131, pp. 193–204, 2020.

[209] D. Lin, Y. Li, T. L. Nwe, S. Dong, and Z. M. Oo, "Refineu-net: Improved u-net with progressive global feedbacks and residual attention guided local refinement for medical image segmentation," *Pattern Recognition Letters*, vol. 138, pp. 267–275, 2020.

[210] S. Dutta, P. Sasmal, M. Bhuyan, and Y. Iwahori, "Automatic segmentation of polyps in endoscopic image using level-set formulation," in *2018 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*. IEEE, 2018, pp. 1–5.

[211] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

[212] L. Zhang and Q. Ji, "Image segmentation with a unified graphical model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 8, pp. 1406–1425, 2010.

[213] W. Feng, J. Jia, and Z.-Q. Liu, "Self-validated labeling of markov random fields for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 10, pp. 1871–1887, 2010.

[214] H. Deng and D. A. Clausi, "Unsupervised image segmentation using a simple mrf model with a new implementation scheme," *Pattern recognition*, vol. 37, no. 12, pp. 2323–2335, 2004.

[215] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," in *Readings in Computer Vision*. Elsevier, 1987, pp. 564–584.

[216] S. Chen, L. Cao, Y. Wang, J. Liu, and X. Tang, "Image segmentation by map-ml estimations," *IEEE Transactions on Image Processing*, vol. 19, no. 9, pp. 2254–2264, 2010.

[217] H. Deng and D. A. Clausi, "Unsupervised image segmentation using a simple mrf model with a new implementation scheme," *Pattern recognition*, vol. 37, no. 12, pp. 2323–2335, 2004.

[218] J. D. Peter, S. L. Fernandes, and C. E. Thomaz, *Advances in Computerized Analysis in Clinical and Medical Imaging*. CRC Press, 2019.

[219] Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm," *IEEE transactions on medical imaging*, vol. 20, no. 1, pp. 45–57, 2001.

[220] J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, "Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer," *International Journal of Computer Assisted Radiology and Surgery*, vol. 9, no. 2, pp. 283–293, 2014.

[221] S. Esedoglu and J. Shen, "Digital inpainting based on the mumford–shah–euler image model," *European Journal of Applied Mathematics*, vol. 13, no. 4, pp. 353–370, 2002.

[222] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual u-net," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 749–753, 2018.

[223] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv preprint arXiv:1412.7062*, 2014.

[224] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, "Pranet: Parallel reverse attention network for polyp segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 263–273.

[225] P. Isola, J.-Y. Zhu, T. Zhou, and A. 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.

[226] F. Deeba, F. M. Bui, and K. A. Wahid, "Computer-aided polyp detection based on image enhancement and saliency-based selection," *Biomedical Signal Processing and Control*, vol. 55, p. 101530, 2020.

[227] H. Ding, Q. Cen, X. Si, Z. Pan, and X. Chen, "Automatic glottis segmentation for laryngeal endoscopic images based on u-net," *Biomedical Signal Processing and Control*, vol. 71, p. 103116, 2022.

[228] J. D. Edwards, K. J. Riley, and J. P. Eakins, "A visual comparison of shape descriptors using multi-dimensional scaling," in *International Conference on Computer Analysis of Images and Patterns*. Springer, 2003, pp. 393–401.

[229] R. S. Montero and E. Bribiesca, "State of the art of compactness and circularity measures," in *International mathematical forum*, vol. 4, no. 27, 2009, pp. 1305–1335.

[230] R. M. Rangayyan, N. M. El-Faramawy, J. L. Desautels, and O. A. Alim, "Measures of acutance and shape for classification of breast tumors," *IEEE Transactions on medical imaging*, vol. 16, no. 6, pp. 799–810, 1997.

[231] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal, "The importance of skip connections in biomedical image segmentation," in *Deep learning and data labeling for medical applications*. Springer, 2016, pp. 179–187.

[232] J. Dolz, I. B. Ayed, and C. Desrosiers, "Unbiased shape compactness for segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 755–763.

[233] Y. Li, K. Zhang, W. Shi, and Z. Jiang, "Lung fields segmentation based on shape compactness in chest x-ray images," *Journal of Computers*, vol. 32, no. 4, pp. 152–165, 2021.

[234] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 9, pp. 1124–1137, 2004.

[235] J. Kim, D. Han, Y.-W. Tai, and J. Kim, "Salient region detection via high-dimensional color transform and local spatial support," *IEEE transactions on image processing*, vol. 25, no. 1, pp. 9–23, 2015.

[236] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Advances in neural information processing systems*, 2007, pp. 545–552.

[237] P. Sasmal, M. Bhuyan, S. Gupta, and Y. Iwahori, "Detection of polyps in colonoscopic videos using saliency map-based modified particle filter," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–9, 2021.

[238] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[239] S. Feng, H. Zhao, F. Shi, X. Cheng, M. Wang, Y. Ma, D. Xiang, W. Zhu, and X. Chen, "Cpfnet: Context pyramid fusion network for medical image segmentation," *IEEE transactions on medical imaging*, vol. 39, no. 10, pp. 3008–3018, 2020.

[240] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.

[241] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine learning*, vol. 3, no. 1, pp. 1–122, 2011.

[242] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE transactions on systems, man, and cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[243] L. M. W. K. Song, D. G. Adler, J. D. Conway, D. L. Diehl, F. A. Farraye, S. V. Kantsevoy, R. Kwon, P. Mamula, B. Rodriguez, R. J. Shah *et al.*, "Narrow band imaging and multiband imaging," *Gastrointestinal endoscopy*, vol. 67, no. 4, pp. 581–589, 2008.

[244] C. Gheorghe, "Narrow-band imaging endoscopy for diagnosis of malignant and premalignant gastrointestinal lesions," *Journal of gastrointestinal and Liver Diseases*, vol. 15, no. 1, p. 77, 2006.

[245] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 2169–2178.

[246] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[247] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1. Ieee, 2005, pp. 886–893.

[248] M. Crosier and L. D. Griffin, "Using basic image features for texture classification," *International journal of computer vision*, vol. 88, no. 3, pp. 447–460, 2010.

[249] M. Varma and A. Zisserman, "A statistical approach to material classification using image patch exemplars," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 11, pp. 2032–2047, 2008.

[250] J.-L. Starck, E. J. Candès, and D. L. Donoho, "The curvelet transform for image denoising," *IEEE Transactions on image processing*, vol. 11, no. 6, pp. 670–684, 2002.

[251] M. N. Do and M. Vetterli, "The contourlet transform: an efficient directional multiresolution image representation," *IEEE Transactions on image processing*, vol. 14, no. 12, pp. 2091–2106, 2005.

[252] W.-Q. Lim, "The discrete shearlet transform: A new directional transform and compactly supported shearlet frames," *IEEE Transactions on Image Processing*, vol. 19, no. 5, pp. 1166–1180, 2010.

[253] K. Falconer, *Fractal geometry: mathematical foundations and applications*. John Wiley & Sons, 2004.

[254] T. Song, H. Li, F. Meng, Q. Wu, and J. Cai, "Letrist: Locally encoded transform feature histogram for rotation-invariant texture classification," *IEEE Transactions on circuits and systems for video technology*, vol. 28, no. 7, pp. 1565–1579, 2017.

[255] S. K. Roy, N. Bhattacharya, B. Chanda, B. B. Chaudhuri, and D. K. Ghosh, "Fwlbp: a scale invariant descriptor for texture classification," *arXiv preprint arXiv:1801.03228*, 2018.

[256] A. P. Pentland, "Fractal-based description of natural scenes," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 6, pp. 661–674, 1984.

[257] B. B. Mandelbrot, *The fractal geometry of nature*. WH freeman New York, 1983, vol. 173.

[258] O. S. Al-Kadi, D. Watson *et al.*, "Texture analysis of aggressive and nonaggressive lung tumor ce ct images," *IEEE transactions on biomedical engineering*, vol. 55, no. 7, pp. 1822–1830, 2008.

[259] N. Sarkar and B. B. Chaudhuri, "An efficient differential box-counting approach to compute fractal dimension of image," *IEEE Transactions on systems, man, and cybernetics*, vol. 24, no. 1, pp. 115–120, 1994.

[260] R. N. Khushaba, S. Kodagoda, S. Lal, and G. Dissanayake, "Driver drowsiness classification using fuzzy wavelet-packet-based feature-extraction algorithm," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 1, pp. 121–131, 2010.

[261] B. Scholkopf and A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

[262] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "Rusboost: A hybrid approach to alleviating class imbalance," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 40, no. 1, pp. 185–197, 2009.

[263] T. Song, J. Feng, L. Luo, C. Gao, and H. Li, "Robust texture description using local grouped order pattern and non-local binary pattern," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 1, pp. 189–202, 2020.

[264] F. Riaz, F. Vilarino, M. D. Ribeiro, and M. Coimbra, "Identifying potentially cancerous tissues in chromoendoscopy images," in *Iberian Conference on Pattern Recognition and Image Analysis*. Springer, 2011, pp. 709–716.

[265] P. Sasmal, M. Bhuyan, Y. Iwahori, and K. Kasugai, "Colonoscopic polyp classification using local shape and texture features," *IEEE Access*, vol. 9, pp. 92 629–92 639, 2021.

[266] G. Koch, R. Zemel, R. Salakhutdinov *et al.*, "Siamese neural networks for one-shot image recognition," in *ICML deep learning workshop*, vol. 2. Lille, 2015.

[267] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[268] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[269] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[270] R. Fonollà, M. Smyl, F. Van Der Sommen, R. M. Schreuder, E. J. Schoon *et al.*, "Triplet network for classification of benign and pre-malignant polyps," in *Medical Imaging 2021: Computer-Aided Diagnosis*, vol. 11597. International Society for Optics and Photonics, 2021, p. 1159731.

[271] K. Bora, M. Bhuyan, K. Kasugai, S. Mallik, and Z. Zhao, "Computational learning of features for automated colonic polyp classification," *Scientific Reports*, vol. 11, no. 1, pp. 1–16, 2021.

[272] R. S. Gonzalez, "Updates and challenges in gastrointestinal pathology," *Surgical Pathology Clinics*, vol. 13, no. 3, p. ix, 2020.

[273] D. P. Kingma, S. Mohamed, D. Jimenez Rezende, and M. Welling, "Semi-supervised learning with deep generative models," *Advances in neural information processing systems*, vol. 27, 2014.

[274] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[275] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[276] Y. Ma, X. Chen, W. Zhu, X. Cheng, D. Xiang, and F. Shi, "Speckle noise reduction in optical coherence tomography images based on edge-sensitive cgan," *Biomedical optics express*, vol. 9, no. 11, pp. 5129–5146, 2018.

[277] J.-Y. Zhu, T. Park, P. Isola, and A. 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.

[278] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690.

[279] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 53–65, 2018.

[280] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[281] P. Costa, A. Galdran, M. I. Meyer, M. Niemeijer, M. Abràmoff, A. M. Mendonça, and A. Campilho, "End-to-end adversarial retinal image synthesis," *IEEE transactions on medical imaging*, vol. 37, no. 3, pp. 781–791, 2017.

[282] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Synthetic data augmentation using gan for improved liver lesion classification," in *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*. IEEE, 2018, pp. 289–293.

[283] A. Odena, "Semi-supervised learning with generative adversarial networks," *arXiv preprint arXiv:1606.01583*, 2016.

[284] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *Advances in neural information processing systems*, vol. 29, pp. 2234–2242, 2016.

[285] V. Kumar, A. K. Abbas, N. Fausto, and J. C. Aster, *Robbins and Cotran pathologic basis of disease, professional edition e-book*. Elsevier health sciences, 2014.

[286] M. Taherian, S. Lotfollahzadeh, P. Daneshpajouhnejad, and K. Arora, "Tubular adenoma," *StatPearls [Internet]*, 2020.

[287] M. Ünlü, E. Uzun, G. Bengi, Ö. Sağol, and S. Sarıoğlu, "Molecular characteristics of colorectal hyperplastic polyp subgroups," *The Turkish Journal of Gastroenterology*, vol. 31, no. 8, p. 573, 2020.

[288] R. K. Pai, M. Bettington, A. Srivastava, and C. Rosty, "An update on the morphology and molecular pathology of serrated colorectal polyps and associated carcinomas," *Modern Pathology*, vol. 32, no. 10, pp. 1390–1415, 2019.