Polyp Segmentation and Identification from Endoscopic Images by Implementing Deep Learning Models

by

Hasin Arman Prokriti
20201092
Tasnia Jannat Ayesha
21301611

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
BRAC University
June 2025

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at BRAC University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Hasin Arman Prokriti	Tasnia Jannat Ayesha
20201092	21301611

Approval

The thesis/project titled "Polyp Segmentation and Identification and from Endoscopic Images by Implementing Deep Learning Models" submitted by

1. Hasin Arman Prokriti(20201092)
2. Tasnia Jannat Ayesha(21301611)

Of Spring, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on June 18, 2025.

Examining Committee:

Supervisor:
(Member)

Dr. Amitabha Chakrabarty
Professor
Department of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam
Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

Colorectal cancer (CRC) is one of the leading causes of cancer-related mortality worldwide, with early detection of colorectal polyps playing a vital role in reducing its impact. Manual identification of polyps during colonoscopy is often limited by human error, variability in polyp appearance, and real-time diagnostic constraints. This work explores a deep learning-based approach to automatic polyp segmentation in endoscopic images, aiming to enhance detection accuracy and reduce diagnostic delays. Multiple convolutional and transformer-based segmentation models are implemented to learn pixel-level features associated with polyp structures. The publicly available Kvasir-SEG dataset is utilized and further augmented to improve training diversity. A specialized boundary-aware loss function is introduced to address the challenge of ambiguous polyp borders. The proposed method emphasizes the importance of robust preprocessing, architectural design, and edge-focused optimization to advance the reliability of computer-aided diagnosis systems in clinical gastroenterology.

Keywords: Polyp Detection, Polyp Segmentation, Deep Learning, Semantic Segmentation, Image Processing, Medical Image Analysis, Machine learning, Data Augmentation, Model Accuracy, U-Net,U-Net++, ResUnet, ResUnet++, DeepLab V3, FPN, Swin-Unet, Automated Diagnosis, Colonoscopy

Acknowledgement

We would like to thank our supervisor, Dr. Amitabha Chakrabarty, and acknowledge his consistent effort in guiding us every step of the way. We would also like to thank Fahim sir for his technical guidance. Additionally, we would like to thank our parents for always supporting us. We are incredibly grateful for everyone's support, mentioned and not mentioned.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Motivation

Colorectal cancer (CRC) is the third most common type of cancer worldwide. Over 1.9 million new cases were recorded and approximately 935,000 deaths in 2020 alone, according to the World Health Organization (WHO) [17]. The prevention and early detection of CRC are critical public health priorities. One of the identifying features of this disease is the build up of polyps in colons and rectums. Colonoscopy is the common technique used to detect the polyps where endoscopic images are analyzed to detect the polyp growth. Currently with advancement in technology Computer Aided Detection (CAD) systems such as deep learning models are being used to help doctors to detect polyps more accurately and in a shorter time. These models can assist in the analysis of endoscopic images to point out the areas of interest that might contain abnormalities When it is applied to detect early polyps there is less chance for human errors during the evaluation step. The deep learning models, especially networking with image segmentation models, have significantly enhanced CAD systems by providing certain reliable tools for the real-time detection, localization and segmentation of polyps during colonoscopy procedures. In this research, we will apply several deep learning techniques for the segmentation evaluation and detection of the polyps from the endoscopic images. The aim is to compare the effectiveness of these models for the polyp detection problem, regarding its specifics which consist of variations in polyp shapes and sizes, and the requirement for fast performance in clinical settings. This report will provide a literature review of the recent development with a focus on how these models can be enhanced for better clinical performance to segment and detect polyp efficiently.

1.2 Research Statement

Segmentation and identification of the polyps in the endoscopic images are still difficult due to the variation in polyp appearance, overlapping between polyp and surrounding tissues and similarity with background structures. The manual detection techniques end up missing a large number of polyps that can cause late detection of colorectal cancer. Furthermore, the lack of availability of labeled medical data sets raises many limitations on how well one can train accurate models. In this research we will focus on the development of a new, more effective, and precise deep learning-based method involving multi-scale feature extraction, attention mechanism, and boundary-aware segmentation to enhance polyp detection and segmentation from endoscopic images in real-time clinical practice settings.

1.3 Research Objectives

Data Collection and Analysis:

- Conduct a comprehensive analysis of endoscopic images collected from public sources such as Kvasir-SEG to identify and segment polyps accurately.
- Extract meaningful features from the images, focusing on the various shapes, sizes, and textures of polyps.

Application of Models:

- Implement advanced deep learning architectures such as U-Net, ResUNet++, and FCN and modify them accordingly for effective polyp segmentation and identification.
- Explore the integration of multi-scale feature extraction and attention mechanisms to enhance segmentation performance.
- Develop a boundary-aware segmentation model that improves the detection of polyp boundaries and minimizes missed detections.

Performance Evaluation:

- Evaluate the performance of the developed models on standard polyp segmentation datasets, utilizing metrics such as Intersection over Union (IoU) and Dice coefficient.
- Benchmark the models against state-of-the-art methods to assess improvements in accuracy, precision, and recall in polyp detection and segmentation.

Chapter 2

Literature Review

2.1 Survey Methodology

For the literature review of our research, we have read many technical research articles and survey papers related to the segmentation of medical images, more importantly polyp segmentation and detection using various deep learning techniques. We focused on choosing relevant papers which had the higher citation counts. We have also tried to include more recent papers as technology is rapidly evolving. However, we have tried to minimize the selection of almost identical papers. We will first discuss generalized papers chronologically which worked on different medical images. Then we will chronologically analyze the papers published dedicatedly working on polyp segmentation and identification.

2.2 Related Works

Ronneberger et al. [1] had introduced the U-Net architecture that was specifically designed to deal with biomedical image segmentation problems, where training data is limited. They aimed to develop a deep learning model which can do segmentation of medical images using the least number of annotated images. The reason is that annotated medical images are very difficult and time consuming to prepare and traditional deep learning models usually need a large amount of data for training. The researchers utilized two main datasets for training and testing: the PhC-U373 dataset consisting of phase contrast microscopy images of U373 cells, and the DIC-HeLa dataset with HeLa cells imaged using differential interference contrast. The PhC-U373 dataset had 35 annotated training images, and the DIC-HeLa dataset had 20 training images.When using the PhC-U373 dataset, U-net achieved an average Intersection over Union (IoU) of 92%. On the DIC-HeLa dataset, U-Net achieved an average IoU of 77.5%. The values were significantly better than the other methods in existence. The U-Net architecture is based on a symmetric design consisting of a contracting path for capturing context and an expanding path for localization. This structure allows the model to learn spatial features well, while keeping fine details. The authors also used data augmentation techniques, such as elastic deformations, by which they increased the variability and robustness of the training dataset. Benefits of U-Net algorithm include high accuracy with small amount of training data, fast processing speed (less than 1 second for segmenting 512x512 image on GPU) and availability of implementation of U-Net algorithm already public. Nevertheless, the architecture is complex, dependent on high quality data and requires computational resources such as GPUs, which may not be possible to some users. Even though

this research paper used microscopic cell images, we can consider implementing this technique due to its benefits mentioned above and modify accordingly to our polyp dataset to segment them.

X. Li et al. [4] introduce H-DenseUNet, a mixed neural network model made to help better identify the liver and tumors in CT scans. This study focuses on the problem of automatically finding the liver and tumors in CT images, which is very important for diagnosing liver diseases and planning treatments such as liver cancer therapy. Regular segmentation methods have difficulty with different tumor sizes, shapes, and positions, leading to results that are not reliable. This work employed data from two sources: MICCAI 2017 Liver Tumor Segmentation Challenge-LiTS, including 131 training and 70 testing CT scans, and the 3DIRCADb Dataset, which contains 20 CT scans. Such datasets provide diverse liver and tumor images and will enable the model to be tested on different cases. A hybrid model that combines both 2D and 3D DenseUNet networks, H-DenseUNet, was used in the study. This combination captures both intra-slice (2D) and inter-slice (3D) spatial features, offering a significant improvement in segmentation accuracy compared to traditional models. The key strength of H-DenseUNet is that it can well balance the accuracy between small and large tumors, which is the greatest value of this approach in clinical applications. However, its complexity and the high computational cost since it combines a 2D and a 3D model are the potential drawbacks. Even with this, the model looks very promising for diagnosing liver cancer and planning treatment. In their study, they found that H-DenseUNet did considerably better: 72.2% Dice score for lesions detection and 96.1% for liver detection. It worked particularly well for large tumors, showing a big improvement compared to other methods.

In 2019, Weng et al. [7] proposed a new way, NAS-Unet, which utilizes Neural Architecture Search (NAS) to automatically design neural network architectures for medical image segmentation, where the objective was to achieve improved segmentation performance with fewer parameters and resolve the problems with the manual architecture design. The focus was on addressing the main challenges of manual design being inefficient, imbalanced medical datasets' performance, and the ability to reduce parameter counts without loss of accuracy. The study utilized three primary datasets: T2-weighted MR prostate images (Promise12, 50 training cases, resized to 256x256), CHAOS Challenge (with CT and MRI of abdominal organs for liver and organ segmentation), Ultrasound Nerve Dataset (for nerve segmentation). In addition, this paper experiments with different dataset than polyp images. Based on the backbone of the U-Net model, NAS-Unet was created by automatically designing the architecture through NAS optimized by two cell architectures, DownSC (Down Sampling Cell) and UpSC (Up Sampling Cell), in order to balance accuracy and parameter efficiency. The acceleration was done with a memory saving binary gate algorithm during the search. The model demonstrated substantial improvements in segmentation performance across datasets:

- Promise12: Dice coefficient of 0.87 (vs. U-Net 0.75).
- CHAOS Challenge: Liver segmentation achieved dice score of 0.85 vs. 0.78 baseline.
- Ultrasound Nerve Segmentation: Dice coefficient of 0.80 (vs. 0.65 for U-Net).

These results were achieved with only 0.8M parameters, much less than U-net's 31M, and shows the parameter efficiency of NAS-Unet. An approach using the NAS, however, provides some benefits, including automatic architecture design, optimization of performance, and robust generalization to different datasets. One drawback is it is computationally expensive, another is that the architectures it creates are more difficult to interpret, and yet another is that it is dependent on data quality.

Hao et al. [9] provide an overview of deep learning methods for semantic segmentation. The paper discusses various datasets commonly used by researchers. These datasets help computers learn how to identify and categorize every aspect of the picture as images are thoroughly labeled for each pixel. In this manner, the models are able to understand what each pixel represents. One dataset is PASCAL VOC, a collection of several image types with labels indicating their contents. Another is Cityscapes, a dataset that specializes in street scenes, with very detailed labels. They primarily focus on deep learning technologies, particularly Fully Convolutional Neural Networks (FCNs). FCNs modify traditional CNNs by replacing fully connected layers with convolutional layers, where each pixel must be classified. FCNs are widely used as they keep the important details of the image and can work on pictures of different sizes. Several extensions of FCNs are also explored, such as U-Net and DeepLab models. U-Net is an encoder-decoder architecture, it works by using an encoder to capture the important features from the image and a decoder to increase the size back to its original, helping to restore any details that were lost when the image was reduced. DeepLab uses Dilated Convolutions to help the model see the image at different levels of detail without making it more complicated. This helps the model identify and handle objects of different sizes better. The paper explains how attention mechanisms help the model focus on the most important parts of the image, making the segmentation more accurate and the use of lightweight networks such as MobileNet for real-time applications. It also talks about special functions, like pixel-wise cross-entropy and Dice loss, that teach the model to segment images properly, even when there are uneven amounts of different objects. Additionally, transfer learning makes an impact on getting improved performance with less labeled data. While these advanced methods offer impressive results, they also present challenges related to computational costs and real-time performance.

Huang et al. [10] introduces UNet 3+, an enhanced architecture for medical image segmentation, particularly targeting challenges in organ segmentation from CT scans. The main objective is to improve segmentation accuracy, especially for organs like the liver and spleen, by addressing common issues such as false positives, where non-organ areas are mistakenly classified as organs. In traditional segmentation methods, the utilization of multi scale features is extremely difficult. They are useful in sensing both fine details (e.g., organ boundaries) and high level semantic information (e.g., organ location). UNet 3+ wishes to integrate these features to better integrate them and improve the overall segmentation performance. The study employed two primary datasets: The ISBI LiTS 2017 dataset for liver segmentation, together with a spleen segmentation dataset, was derived from 131 contrast enhanced 3D abdominal CT scans from a hospital, under ethical approvals. The liver dataset contained 103 volumes of training and 28 volumes of testing and the spleen dataset was 40 training and 9 testing volumes. Each image was cropped to

320x320 pixels with three channels to increase context for segmentation tasks and for preprocessing. The traditional UNet architecture is extended by Full Scale Skip Connections to combine spatial information at low scale with semantics at higher scales. Additionally, the architecture implements Deep Supervision, meaning that each side output is combined with a Hybrid Loss Function that minimizes the error in segmentation across different levels of detail. SGD was used to train the model. They have found that UNet 3+ significantly outperformed state of the art methods by showing results. On the liver dataset it scored at 0.9675 with a 3.4% improvement in segmentation accuracy compared to UNet++, and on the spleen dataset it scored at 0.9620, a 3.8% improvement. Overall, UNet 3+ yielded an average performance improvement of 2.7 points over traditional UNet and 1.6 points over UNet++ and noted additional gains from full scale deep supervision. One of the great things about Unet3+ is that it's both accurate and efficient, with many key benefits including high accuracy, effective integration of multi-scale features, low parameter count. However, the model's complexity, reliance on high-quality data, and potentially longer training times may pose challenges in clinical or resource-limited settings.

Wang et al. [15] concentrate on the improvement of medical images segmentation, a process that involves recognizing and marking out the structures of the body present in medical images such as MRI, CT scans, and X-rays. This enhancement is important for more reliable computer-aided diagnosis and efficient treatment planning. The authors utilized publicly available medical imaging datasets that the researchers collected from well-known medical imaging databases. These datasets consist of diverse imaging modalities to facilitate investigation across multiple diseases. The research in focus utilizes deep learning tools and techniques especially, Convolutional Neural Networks (CNNs) applied mainly for image processing tasks. The architecture of concern is U-Net, a type of CNN aimed to enhance segmentation by encoders and decoders with skip connections that maintain both low and high-level features of the image. The authors assert that deep learning makes a great leap in the effectiveness and efficiency of medical imaging segmentation processes as compared to previous techniques. These improvements solve important issues such as image noise and blur making diagnoses better and aiding in the production of better medical equipment. However, the costs of computational power and access to plentiful well-annotated datasets constitutes a hindrance to extensive use.

This paper by Cao et al. [16] focuses on the lack of long-range contextual information in medical image segmentation primarily due to the strictly localized process of convolution in CNNs. New for the study, the researchers present Swin-Unet, which is designed to combine the Swin Transformer with an encoder-decoder U-shape architecture to achieve better and more global and local feature learning for increased segmentation precision. The work is based on datasets of multi-organ and cardiac segmentation, which are essential for diagnostics. Although the sources of the datasets aren't detailed, they are likely derived from common medical imaging repositories. Swin-Unet uses tokenized image patches that are then handled with Swin Transformer blocks to capture hierarchical features and to enable the model's engagement with semantics that extend across a wide range of areas. Skip connections also help in increasing the learning of local and global features. The model

gives better performance when compared with conventional CNN-based solutions especially on multi-organ and cardiac segmentation and also demonstrated good ability to generalize. The researchers also propose that in further studies, Swin-Unet can be used to segment 3D medical images as well. Swin-Unet has better algorithmic improvement; feature learning is more accurate; and Swin-Unet transfers or adapts well to other medical imaging tasks. However, the model does present some difficulties: specificity on pre-training, where ImageNet weights are not suitable for medical practice. If the results are compared to previous methods it takes more resources in the form of computational power.

The work done by Brandao et al. [2] proposes a technique to enhance the detection and segmentation of polyps in colonoscopy images in Fully Convolutional Networks (FCNs). This paper and the following papers used tiny images of polyps as their dataset. The aim of the present work is to minimize the miss rate of polyps during colonoscopy in order to improve the chances of early diagnosis of colorectal cancer. Polyps that might be overlooked can lead to delayed diagnosis and hence delayed treatment The proposed system therefore makes a timely diagnosis possible and assists clinicians in identifying polyps more easily and accurately. This enhancement of accuracy in clinical practice of colonoscopy exams is aimed at increasing the general success rates of the tests. The information for the study is derived from publicly available datasets of the MICCAI 2015 polyp detection challenge. Specifically, the first authors collected the images from three datasets: CVC-CLINIC, ETIS-Larib, and ASU-Mayo datasets containing 612 SD images, 196 HD images and both SD and HD images respectively. These various datasets mean better training and testing of polyp detection system; the system's robustness when it comes to interpreted image qualities and conditions. Towards that end, the researchers used Fully Convolutional Networks (FCNs), a sub-type of Convolutional Neural Networks (CNNs) for pixel-level image categorization task. Three classic CNN architectures were firstly established, and then were adjusted and improved to form three novel FCNs for the purpose of polyp segmentation of colonoscopy images. With the proposed FCN-VGG architecture, the detection precision was 73.61% and the recall was 86.31%, which outperforms prior work or evolves from 2015 MICCAI challenge. The conclusion of the study is that anatomical FCNs allow achieving a high accuracy of polyp detection and segmentation, which makes this tool useful in clinical practice. However, the complexity of FCN architectures can result in longer processing times and higher computational demands, which could pose challenges for widespread clinical implementation.

In their study, Jha et al. [6] proposed ResUNet++ for medical image segmentation to enhance the detection and segmentation of polyps from colonoscopy images. The main aim of the proposed research was due to existing problems of the segmentation model such as the cases where different polyps look similar or same type of polyps may look different in various cases. The ResUNet++ architecture proposed in this paper attempts to achieve a higher level of segmentation accuracy, which is helpful to the early diagnosis of colorectal cancer and decision of endoscopists. To train and validate the ResUNet++ model, the authors utilized two key publicly available datasets: Kvasir-SEG with 1000 images marked as polyps and CVC-ClinicDB database which comprises of 612 images from 31 colonoscopy sequences.

These datasets include various orientations and non gastrointestinal objects and backgrounds to enrich the model of polyp segmentation. The research team used the Keras framework with the TensorFlow environment and performed the experiments in an Nvidia DGX-2 AI system with 16 Volta GPUs for high computational power in training the complex architecture. ResUNet++ uses several cutting-edge approaches in order to enhance the segmentation outcome. For smoothing the gradients flow, Residual blocks have been implemented and Squeeze and Excitation blocks for better feature extraction. Spatial pyramid pooling was adopted in the form of Atrous Spatial Pyramid Pooling (ASPP) to capture multi-scale features from the images and attention blocks were used to guide within the feature maps. It was apparent that the model received enhancements to the segmentation points; on the Kvasir-SEG dataset, it earned a Dice coefficient of 0.85, with top rankings over U-net that got a coefficient of 0.78 and ResUNet got a coefficient of 0.80. However, fast convergence achieved by ResUNet++ provides improved accuracy and model robustness, but it has limitations: The increased computational step is required and risks of overfitting if train data is small.

In another research paper by Jha et al. [13] aims at training a deep learning based model for polyp detection, localization and segmentation in colonoscopy images to meet the significant clinical requirement of early identification of colorectal cancer. Conventionally, these tasks are accomplished manually by gastroenterologists, but manual handling makes the performance error-prone and inconsistent. The proposed solution named ColonSegNet aims for more efficient and accurate detection and segmentation of polyps, giving an opportunity to improve clinicians' practice. One of the major objectives of the study is to propose an end-to-end reference which could be used to assess deep learning methods in polyp segmentation since such reference frame is insufficiently developed in the current state. The authors employed the Kvasir-SEG dataset consisting of 1,000 images manually labelled and collected for polyp segmentation purposes. These images show different sizes, shapes, and textures of the polyps, proving the effectiveness of the proposed model in training and validation. The authors augmented data by flipping, rotating, and scaling, which makes ColonSegNet more invariant to variations in polyp appearance or imaging conditions. Like all its counterparts, ColonSegNet includes the strategy of the convolutional neural network which allows recognizing colonoscopy images in comparatively short time—180 frames in the detection mode and 182.38 frames in the segmentation mode. The model showed high performance rates, namely, the average precision constituted 0.85, while the Intersection over Union (IoU) of the segmentation reached 0.78. However, the study has some limitations; for instance, performance varies with polyp characteristics, and additional experiments are required with other databases than Kvasir-SEG. Nevertheless, the present work identified ColonSegNet as a valuable tool for real-time polyp detection into clinical application to increase speed, accuracy, and robustness in the future.

In the work of Fan et al. [8] a new method is presented to enhance the polyp segmentation in colonoscopy images, which is critical for initial-stage colorectal cancer screening. In this regard, problems arise with the variability of sizes, texture, as well as the uncertainty with boundaries of the polyps within the existing traditional segmentation models. To address these challenges, PraNet proposed to utilize a

Parallel Reverse Attention Network that improves the communication between the polyp's area and boundary. It is named this for pointing back to boundaries and as a mechanism that lets the skin area of the polyps give its segmentation greater precision. The authors collected colonoscopy images, and particularly CVC-612 dataset on this work, with five more difficult datasets to make the method more stable in different appearances of polyps. The current design of PraNet is based on the CNNs and it has incorporated the recurrent cooperation mechanism in its attempt to extract features at the area and boundary levels. A comparison with the best existing methods on three benchmark datasets showed that PraNet resulted in an advantage of over 5% for polyp segmentation. This approach showed a high mean Dice coefficient of 0.898 regarding the Kvasir dataset and in real-time medical application, in running approximately at 50 fps. The model also performs well across different datasets since the flexibly chosen r is robust enough to accept changes between different polyp variations. However, the authors note that PraNet might require slightly more computational resources than the other models in this field, and that its effectiveness depends on the quality of the training data considerably. Compared to the previous methods, PraNet provides a solution for accurate and real-time polyp segmentation, yet a further study on generalizing the performance to other unseen data is still the future direction.

L. F. Sanchez-Peralta et al. [11] concerned with the better detection of colorectal polyps during colonoscopy using deep learning methods. Colorectal cancer can be avoided with the help of very early polyp diagnosis. By using CAD systems to help gastroenterologists perform colonoscopy operations, the authors increase the adenoma detection rate in this regard and improve the sensitivity and efficiency of detection. Deep learning techniques on colonoscopy pictures are combined in this review for three main tasks: segmentation, which marks the boundaries of the polyps; localization, which determines their location; and detection, which locates the polyps. The study made use of publicly accessible datasets from the ASU-Mayo Clinic, CVC-ClinicDB, and CVC-ColonDB. These datasets are sets of labeled photos or videos of polyps that are used to perform the required main three tasks. The primary technique used is deep learning, particularly CNNs. CNNs automatically recognize patterns and features in pictures. They specialize at tasks like identifying, tracking down, and separating polyps in recordings of colonoscopies. It increases the polyp identification rate. Other networks, such as fully convolutional networks, carry out the same functions and offer comprehensive polyp outlines. High accuracy and automatic learning of complex picture characteristics are two benefits of CNN algorithms. These models may overfit if not constantly monitored and require large amounts of labeled training data. The study emphasizes the need for testing on bigger, diverse sets and in actual clinics. The techniques that are being offered would be far more beneficial as a result. The authors suggest a format for standardized testing that would make it easier to compare various methods.

Ali et al. [12] strive to enhance the accuracy and reliability of artifact and disease detection, segmentation in GI endoscopy images. These difficulties affect the quality of clinical diagnoses, particularly in those with cancer and precancerous lesions. They used a large, multicenter dataset from both the upper and lower GI tract (gastroscopy and colonoscopy). It was collected from different hospitals compris-

ing a good diversity in terms of the number and type of case allowing analysis on over 280 patient videos, adding up to more than 45,478 annotations. They used state-of-the-art deep learning models like YOLOv3, RetinaNet and Faster R-CNN along with ensemble approaches. These techniques have been applied to both object detection and segmentation, leading to a better performance on various datasets. Ensemble methods not only increases the probability of detection but also utilizes them to use a model by combining strengths. However, models often suffer from artifacts and variations in the data which could lead to errors when applied in a real clinical setting. Although the study has some limitations, it demonstrated that DL methods enhance accuracy for artifact and disease detection in endoscopic images.

Safarov et al. [14] introduces A-DenseUNet in their paper which is a fully automated model designed to segment colorectal polyps from colonoscopy images, aiming to improve early detection of colorectal cancer. Proper segmentation of polyps is crucial since colon cancer is one of the leading killers among cancer diseases. In the model it is used for polyp detection and to separate it from surrounding tissue in endoscopic images to benefit endoscopists. The researchers used two key datasets for model training and testing: Kvasir-SEG dataset with 1000 images and CVC-612 with 612 images. Such dataset offer the variability of appearances of polyps and are acquired from actual clinical scenarios. Images' sizes and resolution were different, and they were all resized to a 224×224 matrix so that the model takes in standardized input data. A-DenseUNet uses Complex state-of-art deep learning approaches to improve its functionality. It applies attention layers to address web-image features inattention and apply dense blocks to optimize features' reuse and gradient flow and apply residual blocks to prevent vanishing gradient. Further, the operation of atrous convolution can better perceive the large scope without losing the high-resolution spatial, a way to increase the degree of accuracy. On the Kvasir-SEG dataset, the model received the Dice coefficient of 0.8912 and an IoU of 0.8553; thus, the proposed model is more efficient in comparison with UNet++, RFCN, and ResUNet. Nevertheless, the proposed A-DenseUNet shows high accuracy and robustness in comparison with other similar networks and it has some drawbacks which consist in a higher number of parameters and potentially longer training time. Another weakness is that the model's data sample is limited, and therefore we might overemphasize its results; hyperparameters could therefore be tuned for the best results. However, the difficulties encountered above make A-DenseUNet a prospective approach for polyp segmentation and will be possible to use in daily practice.

2.2.1 Classical CNN-Based Segmentation Approaches

The seminal architecture **U-Net**, introduced by Ronneberger et al. in 2015 [1], is a foundational model in biomedical image segmentation. Its encoder-decoder structure with symmetric skip connections enables the model to capture both semantic context and high-resolution spatial information. U-Net performs well with limited data and has become the baseline model for numerous medical segmentation tasks. However, its performance degrades when dealing with polyps with faint boundaries, low contrast, or irregular shapes.
To improve training dynamics and deeper context extraction, **ResUNet** integrates

residual learning blocks from ResNet. These blocks help mitigate the vanishing gradient problem in deeper networks, allowing better gradient flow and faster convergence. While ResUNet enhances feature reuse and robustness, it still relies on the U-Net's basic skip-connection design and often struggles with fine-grain boundary reconstruction.

U-Net++ [5] advances the vanilla U-Net by introducing nested and dense skip pathways. This creates a more gradient-resilient structure and supports multi-depth supervision, meaning the decoder learns more detailed information at different abstraction levels. U-Net++ is especially beneficial in scenarios with complex topologies—such as endoscopic polyps—by enhancing boundary localization and multi-scale feature representation. It has demonstrated superior accuracy across several biomedical benchmarks including Kvasir-SEG and CVC datasets.

2.2.2 Dilated Convolution and Multi-Scale Feature Models

DeepLabV3 [3], developed by Google Research, utilizes dilated (atrous) convolutions and Atrous Spatial Pyramid Pooling (ASPP) to achieve better contextual aggregation at multiple scales. This is particularly important in polyp segmentation, where polyps vary greatly in shape and size. ASPP enables capturing long-range dependencies without increasing computation. While DeepLabV3 achieves strong global context modeling, it often lacks precise boundary refinement due to its downsampling-heavy backbone.

Feature Pyramid Network (FPN) presents a complementary approach. Originally designed for object detection, FPN leverages a top-down feature refinement strategy with lateral connections, aggregating both low- and high-level features. Although less common in polyp segmentation tasks, FPN has potential due to its capacity for multi-scale learning. It performs well in detecting polyps with unclear morphology, but may underperform in real-time applications due to inference latency and limited decoder expressiveness.

2.2.3 Transformer-Based Segmentation

With the advent of Vision Transformers (ViT), the medical imaging community has adopted attention mechanisms for long-range context modeling. **Swin-Unet V2** [16] builds upon Swin Transformer blocks and introduces shifted window-based self-attention that captures both local and global dependencies efficiently. This architecture maintains hierarchical representations while avoiding the heavy computation of full self-attention.

Swin-Unet V2 excels at segmenting polyps with heterogeneous textures and occlusions, outperforming traditional CNNs in scenarios requiring high global coherence. However, its adoption in clinical settings is still emerging due to high memory demands, long training times, and sensitivity to hyperparameter tuning. Additionally, its performance depends heavily on large, diverse datasets—limiting its utility on smaller medical datasets unless pretrained weights or self-supervised pretraining is used.

2.2.4 Polyp-Specific and Lightweight Innovations

Several models have been proposed specifically for colorectal polyp segmentation, introducing domain-driven innovations:
PraNet [8], proposed by Fan et al., utilizes reverse attention modules and edge guidance mechanisms to refine polyp boundaries. PraNet performs exceptionally well on small and camouflaged polyps, outperforming earlier models with minimal post-processing. Its cascade decoder and contextual attention modules help highlight weak boundaries missed by conventional models.
ColonSegNet [13] offers a lightweight CNN-based architecture optimized for real-time deployment in clinical environments. It integrates residual modules, edge-attention modules, and early exit strategies to balance accuracy and speed. Although less accurate than PraNet or U-Net++, it is well-suited for mobile or embedded diagnostic systems.
ResUNet++ [6] combines residual blocks, ASPP, and squeeze-and-excitation (SE) modules to enhance both context and focus. It significantly improves model attention to salient foreground features and demonstrates strong generalizability across different datasets such as Kvasir-SEG, ETIS-LaribPolypDB, and CVC-ClinicDB. ResUNet++ is recognized for achieving Dice scores greater than 0.91, especially with carefully tuned loss functions (e.g., Dice + Focal Loss).

The following table 2.1 summarizes the outputs of each paper:

No	Model	Dataset(s)	Key Contributions	Performance	Limitations
[1]	U-Net	PhC-U373	Encoder-decoder architecture with skip connections for medical image segmentation.	Improved segmentation performance.	Requires large annotated datasets and is computationally expensive.
[2]	FCN	CVC-CLINIC, ETIS-Larib, ASU-Mayo	Enhanced accuracy in polyp segmentation for colonoscopy images.	Precision: 73.61%, Recall: 86.31%.	High processing time and computational demand.
[6]	ResUNet++	Kvasir-SEG, CVC-ClinicDB	Combines residual blocks, ASPP, and attention for better handling of polyp variability.	Dice: 0.85.	Computationally expensive; risk of overfitting.
[8]	PraNet	CVC-612 + others	Introduced Parallel Reverse Attention Network for boundary-aware segmentation.	Dice: 0.898 (Kvasir), 50 FPS.	Requires high-quality data; more computational resources.
[13]	ColonSegNet	Kvasir-SEG	Real-time polyp detection and segmentation with benchmark framework.	Precision: 0.85; IoU: 0.78; 180–182 FPS.	Limited dataset size; variation in polyp types affects performance.
[11]	CNN, FCN	ASU-Mayo, CVC-ClinicDB, ColonDB	Reviewed segmentation and detection in CAD systems for endoscopy.	Improved polyp identification rate.	Needs validation on large-scale clinical data.
[12]	YOLOv3 + Ensembles	GI dataset (280 videos)	Used ensemble of detection models for robustness.	Better detection across multiple datasets.	Sensitive to data shifts; clinical deployment challenges.
[14]	A-DenseUNet	Kvasir-SEG, CVC-612	Utilized attention, dense and residual blocks with atrous convolutions.	Dice: 0.8912; IoU: 0.8553.	Long training; high parameter count; generalization limits.
[15]	U-Net	Multiple modalities	Applied U-Net to various medical imaging types.	Enhanced segmentation effectiveness.	High computational requirements; needs large annotated datasets.
[16]	Swin-Unet	Multi-organ, cardiac datasets	Merged Swin Transformer with U-Net for global and local features.	Outperformed CNN-based methods.	Requires specific pretraining; compute-intensive.
[8]	PraNet	Kvasir, CVC-ClinicDB	Boundary-aware segmentation via Parallel Reverse Attention.	Dice: 0.899; 50 FPS.	High model complexity; quality data needed.
[12]	MultiResUNet	Kvasir-SEG, CVC-ClinicDB	Multi-scale feature extraction with MultiRes blocks and residual paths.	Dice: 0.923 (Kvasir), 0.895 (CVC).	Sensitive to hyperparameters; high computation.

Table 2.1: Comparison of related works in polyp segmentation models.

2.3 Research Gaps

- **High Computational Complexity**
  Deep learning models such as U-Net, H-DenseUNet, ResUNet++, and PraNet require high computational power and memory, which limits their use in resource-constrained clinical settings. Real-time performance remains an issue due to increased processing time and long training durations, especially with models like NAS-Unet and UNet 3+.

- **Limited Generalizability Across Datasets**
  Several models, including Swin-Unet and CNN-based U-Net, show varying performance when applied to different datasets. This raises concerns about their ability to generalize to unseen data or real clinical scenarios. Overfitting is also evident in models like ResUNet++ and A-DenseUNet, which struggle with small or limited datasets.

- **Dependence on Large Annotated Datasets**
  Many segmentation models, such as UNet 3+ andPraNet, rely heavily on large, high-quality, manually annotated datasets for training. However, obtaining such data in the medical field is challenging, time-consuming, and costly, restricting model scalability and wider adoption.

- **Usability and Real-Time Deployment Limitations**
  Complex architectures like UNet 3+ and Swin-Unet require significant computational resources and extended training times, making them less suitable for real-time clinical applications. Additionally, these models demand specialized hardware, limiting their accessibility to a broader user base.

- **Optimization and Efficiency Challenges**
  Many models, including ResUNet++, FCN-based models, and H-DenseUNet, struggle to maintain a balance between segmentation accuracy and processing efficiency. Noise and artifact handling, as noted in works like Ali et al. [12], remains inadequate, affecting the robustness of these models in practical medical imaging scenarios.

- **Limited Focus on Boundary Precision**
  Many polyp segmentation models focus on overall overlap metrics (Dice, IoU), but ignore boundary-level precision, which is clinically crucial—especially for flat, irregular, or inflamed polyps where wrong segmentation can mislead diagnosis.

- **Under-utilization of Ensemble and Hybrid Techniques**
  Ensemble and hybrid modeling techniques, which could enhance robustness and generalization, are rarely applied. Models from Ali et al. [12] and Sanchez-Peralta et al. [11] indicate that such strategies could improve performance but have not been fully explored.

Chapter 3

Work Plan

3.1 Workflow

We will complete our thesis within the span of a year. Initially, we collected multiple scholarly articles, survey, and review articles from various sources like researchgate, pubmed, IEEE explore etc.. After gathering these, we have studied and analyzed the documents to look for informations that will aid in our research progress. The final task of phase 1 was to write a literature review which will summarize all these resource materials. In pre-thesis 2 we have gathered KvasirSEG dataset which is a collection of endoscopic images and their corresponding masks. We trained existing segmentation models and evaluated their performances. The final phase of the thesis involves analyzing the results, completing the report, and preparing for the final defense. Here we try to compare and further improve our model to obtain the optimal result.

To address the identified limitations, we followed this experimental work plan as explained below:

- **Multi-Model Benchmarking Framework:**
  We implemented seven segmentation architectures—U-Net, U-Net++, ResUNet, ResUNet++, DeepLabV3, Feature Pyramid Network (FPN), and Swin-Unet V2—under a standardized training pipeline. Each model was trained and tested using identical preprocessing, augmentation, and evaluation metrics to ensure fair and consistent benchmarking.
- **Dataset Processing and Augmentation:**
  We utilized the Kvasir-SEG dataset and appliled various augmentations.
- **Transformer Inclusion—Swin-Unet V2:**
  Swin-Unet V2 was included as a transformer-based baseline. We evaluated its performance relative to CNN counterparts and analyzed its training cost and feasibility in real-world clinical settings.
- **Model Comparison Based on Performance and Complexity:**
  After the default implementation of the models we used various performance metrics to compare and analyse the models.
- **Implementing customized loss function**
  After analyzing performances of default model setups, we implemented our customized loss function to enhance the models further

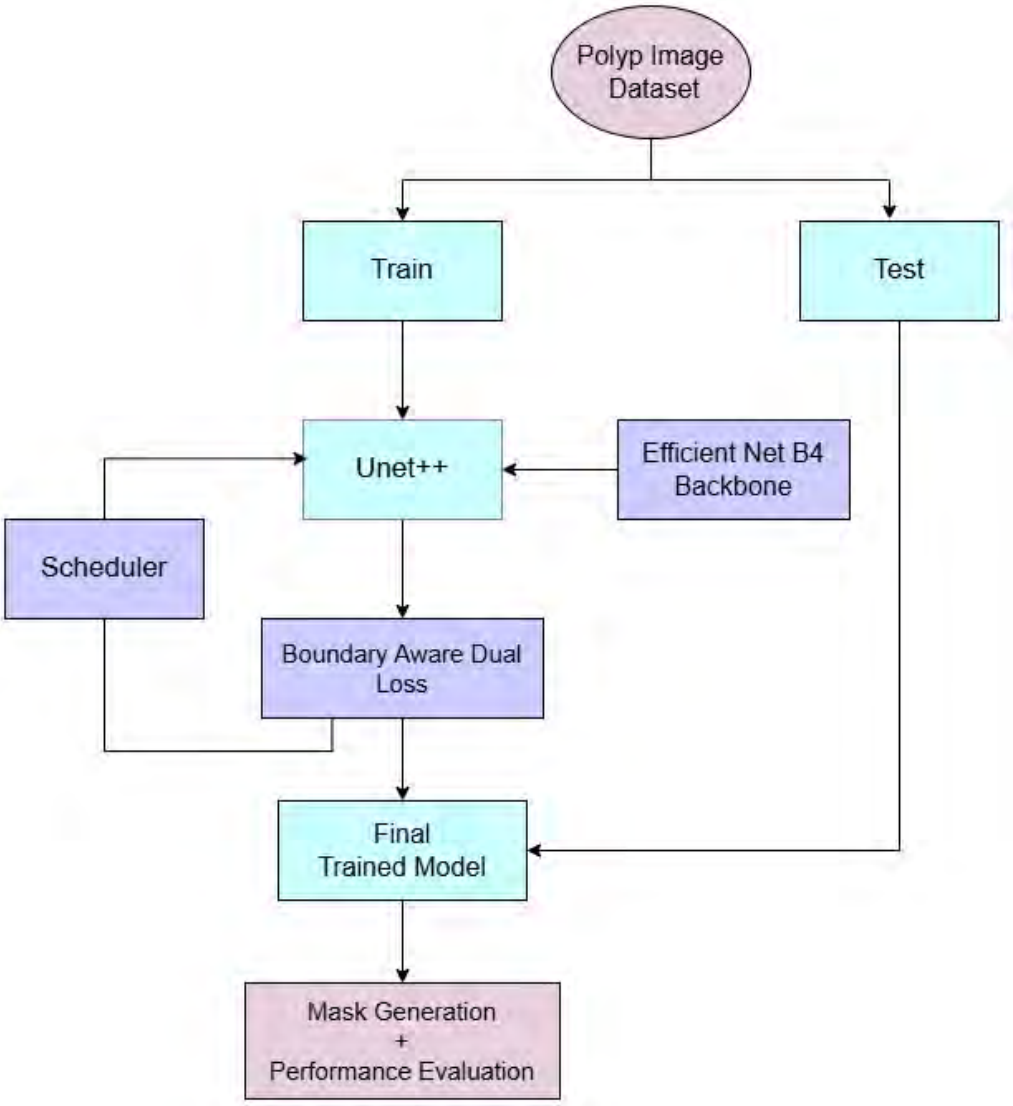

Figure 3.1: Proposed Syestem

3.2 Proposed Model

After implementing and analyzing various deep learning models, we fine tuned and implemented our proposed **Bounday Aware Dual Loss** function on our best performing model. We elaborate our findings and reasoning behind choosing this function in **Chapter 4** and **Chapter 5**. The following Figure 3.1 is the top-level overview of our proposed system.

Chapter 4

Methodology and Implementation

4.1 Dataset

For our research we have chosen the Kvasir-SEG dataset; which is an open-source dataset.This dataset contains 1000 polyp images and their corresponding ground truth from the Kvasir Dataset v2. Due to our lack of medical experties we have opted for globally recognised dataset for polyps as the masks are identified and segmented by specialists from Oslo University Hospital in Norway [6]. The resolution of the images contained in Kvasir-SEG varies from 332x487 to 1920x1072 pixels. The images and its corresponding masks are stored in two separate folders with the same filename. The image files are encoded using JPEG compression, and online browsing is facilitated. However, because the quantity of this dataset is low, we relied on augmentation techniques to increase the quantity of the dataset. Figure 4.1 shows an image and its corresponding mask from the Kvasir Dataset.

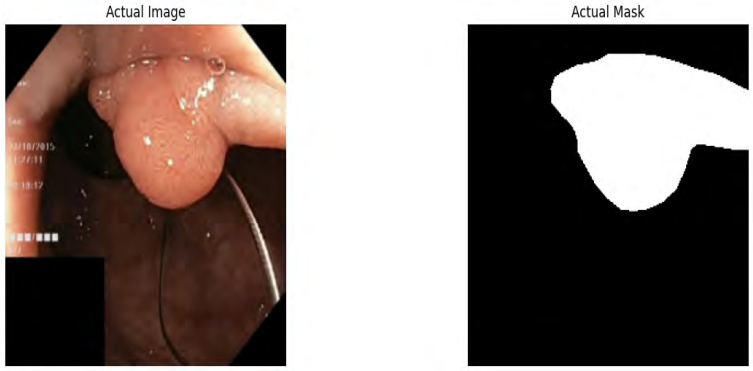

Figure 4.1: KvasirSEG dataset sample image

4.1.1 Data Preprocessing

The Kvasir-SEG dataset v2, comprising 1,000 polyp images and their corresponding ground truth masks, was utilized for this study. The dataset contains images with resolutions ranging from 332×487 to 1920×1072 pixels, stored in separate folders for images and masks, with identical filenames for easy pairing. To prepare the dataset for training and evaluation, all images were loaded and converted to three-channel RGB format, while masks were converted to grayscale and binarized, ensuring pixel values were either 0 or 1 to suit binary segmentation tasks. Both images and masks were resized to a fixed resolution of 256×256 pixels, maintaining consistent input dimensions for the deep learning models.

4.1.2 Data Augmentation

In medical image segmentation tasks such as polyp detection in endoscopic images, data augmentation is essential due to the relatively small dataset sizes and high variability in tissue appearance. The following augmentations were chosen to improve the robustness and generalization of the model while preserving anatomical structures:

- **Horizontal Flip**: The horizontal flip transformation reverses the image and its corresponding mask from left to right. This is particularly helpful in polyp segmentation, as polyps can appear on either side of the colon. By flipping the images, the model learns to detect polyps regardless of their horizontal location. This augmentation is safe and widely used, applied with a probability of 50% during training.

- **Vertical Flip**: Vertical flipping mirrors the image along the vertical axis. While this transformation is less common in endoscopic imaging—since the orientation of the colon is generally consistent—it can still introduce helpful variation. However, excessive use may introduce unrealistic views, so it is applied cautiously, with a low probability of 20%.

- **Rotation**: Rotation involves turning the image and mask by a random degree within a predefined range, typically ± 30 degrees. This augmentation accounts for slight tilts or shifts in camera angle that naturally occur during endoscopic procedures. Applying moderate rotations allows the model to recognize polyps in varied orientations without distorting the anatomical structure. This is implemented with a probability of about 60%.

- **Random Resized Crop**: This transformation randomly crops a portion of the image and then resizes it to the original dimensions. It helps the model focus on different parts of the image and learn scale-invariant features, which is important since polyps may vary in size and location. By introducing slight zooming in and out, this augmentation enhances the model's flexibility, and it is applied in about 50% of the training samples.

- **Elastic Transform**: Elastic deformation slightly warps the image in a manner that mimics the flexible nature of human tissue. This augmentation simulates realistic anatomical variations due to stretching and compression, which often occur during scope navigation. It encourages the model to learn robust boundary features. However, since excessive distortion can lead to unrealistic examples, it is used sparingly with a 20–30% chance.

- **Grid Distortion**: Grid distortion applies a gentle, grid-like warping to the image. This can simulate the minor optical distortions introduced by the endoscope lens or camera misalignments. It helps the model generalize to slightly distorted views of the anatomy. As with elastic transform, its use is conservative, with around 20% probability, to maintain realistic visual representations.

- **Brightness and Contrast Adjustment**: This augmentation alters the brightness and contrast of the image randomly within a realistic range. Variations in lighting are common in endoscopic procedures due to different angles,

equipment, and depths. Modifying these properties helps the model adapt to such inconsistencies. It is applied in 50% of the training images with controlled intensity, avoiding unnatural illumination.

- **Hue, Saturation, and Value (HSV) Shift**: Adjusting hue, saturation, and value components introduces slight changes in the color appearance of the image. These changes reflect real-world variations caused by different lighting environments or equipment settings. In polyp segmentation, where color and texture are both important, this augmentation helps the model learn to rely on more generalizable features. Since significant changes could introduce unrealistic tones, it is applied with a low probability of 30%.

- **Normalization**: Normalization is a standard preprocessing step where pixel values are scaled based on the dataset's mean and standard deviation. This ensures that the input distribution is consistent across all samples, which improves training stability and speeds up convergence. Normalization does not alter the semantic content of the image and is always applied as the final step before feeding data into the model.

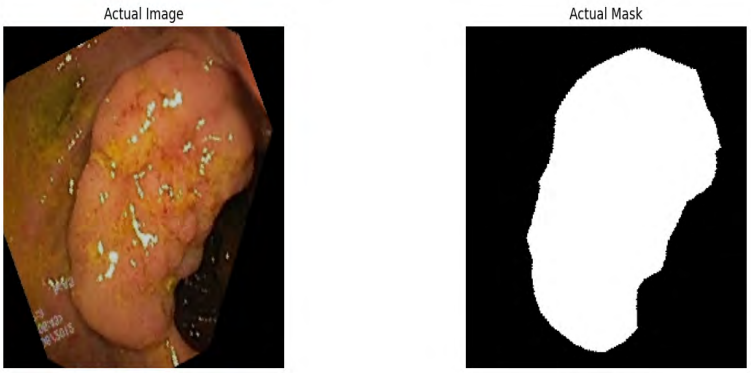

Figure 4.2: KvasirSEG dataset augmented image

We can see in figure 4.2 that the image is left rotated and the lighting is also been changed to augment the image.

4.1.3 Data Splitting

The dataset was then divided into two distinct subsets, both for training and evaluation of the model. First, the combined dataset, comprising a total of 2,000 images with both original and augmented data, was divided into 80% for training and 20% for testing. The division ensured that the model has a good number of samples to learn from and also kept images for unbiased evaluation. testing. Precisely, the training set contains 1,600 images, while the remaining 400 were left for testing. Whereas the training set is helpful in optimizing the model parameters, the testing set will be fair to estimate generalization capability. This is the standard ratio, which gives a good balance: sufficient data for training with success but still keeping a reasonable test set on which to compute the different performance metrics like accuracy, precision, and recall.

4.2 Experimental Setup and configuration

We have run existing models with the following specifications: Intel Core i7-14700k CPU with base 3.4GHz, 64GB DDR4 RAM, and NVIDIA RTX 4080 Super with 16GB VRAM. This amount of hardware resource could provide adequate computational resources and a smooth run in training or rechecking performance using deep learning models. It has used the system PyTorch cuda version 12.1, one of the most popular deep learning frameworks, along with CUDA for GPU acceleration and the python version was 3.9.21. The purpose of CUDA enabling was to utilize the high speed matrix calculation capability of the GPUs which significantly reduces training time compared to CPU-based processing.

We used the "segmentation-models-pytorch" framework to initialize the models, which provides cutting-edge architectures for semantic segmentation tasks. We have used 7 segmentation models from this repository, and the same training conditions were followed for all the models. The ResNet-34 encoder was used with pre-trained ImageNet weights, which would allow the model to utilize transfer learning. The model was configured with 3 input channels (RGB images) and 1 output channel for binary segmentation. The training loop ran for 20 epochs with the learning rate set to 1e-4 for the Adam optimizer, chosen because of its good efficiency in optimizing deep learning models. The loss function selected was Binary Cross-Entropy with Logits Loss (BCEWithLogitsLoss), as it suits binary segmentation tasks well. Initialization and training on the GPU were performed through PyTorch's method, where the was obtained based on CUDA availability of course, to ensure optimal performance throughout the experiment.

4.3 Models

4.3.1 Unet

The U-Net architecture consists of a symmetrical encoder-decoder structure. The encoder, also called the contracting path, progressively reduces the spatial dimensions of the input image through a series of convolutional and max-pooling layers. This path is responsible for extracting increasingly abstract and high-level features. On the other side, the decoder—or expanding path—gradually upsamples the feature maps to restore the original input size. Crucially, skip connections are used to directly transfer feature maps from the encoder to the decoder at corresponding levels, preserving spatial context and fine details that might be lost during downsampling.

In our setup, each input polyp image (RGB, 256×256) enters the U-Net model and passes through successive encoder layers, where it is convolved and downsampled, transforming from a 3-channel image to deeper feature maps with increasing channels and reduced spatial dimensions—such as from 256×256×3 to 128×128×64, then down to 64×64×128, and so on. At the bottleneck, the deepest layer captures the most abstract representation of the image. Then, in the decoder, these features are upsampled and concatenated with the corresponding encoder outputs via skip connections. This process helps refine the segmentation output while preserving

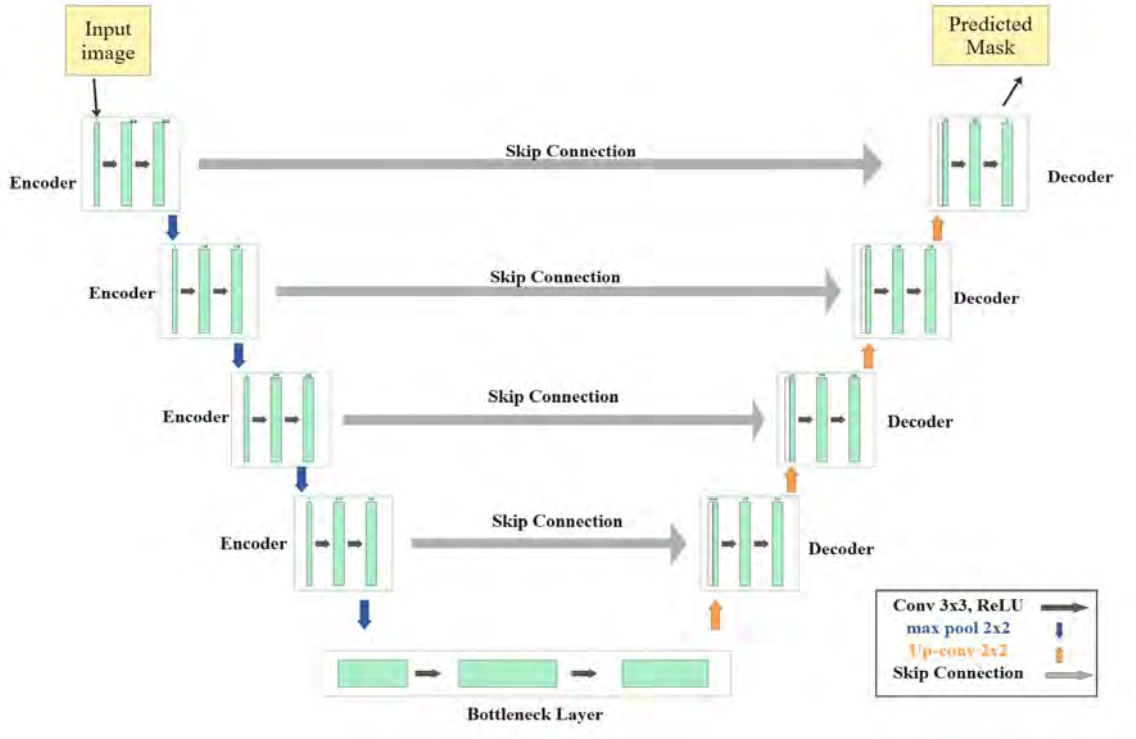

Figure 4.3: Unet model architecture

structural accuracy. Finally, the model outputs a single-channel prediction map of size 256×256, where each pixel value represents the predicted probability of that pixel being part of the polyp region. This output is then thresholded to generate the final binary segmentation mask.

4.3.2 Unet++

The U-Net++ model is a refined extension of the traditional U-Net architecture, specifically designed to address the semantic gap between encoder and decoder feature maps. It introduces nested skip connections and deep supervision, where the decoder is constructed with a series of intermediate convolutional blocks that connect encoder and decoder layers more densely. These additional paths allow the model to gradually bridge the semantic differences between the features extracted at different depths, leading to richer feature representations and more accurate learning of complex spatial details. The design makes U-Net++ more capable of handling intricate structures in medical segmentation tasks compared to the original U-Net.

In our implementation setup, both U-Net and U-Net++ process the Kvasir-SEG polyp images (RGB) and corresponding binary masks through a similar encoder-decoder framework with a ResNet-34 backbone. However, in U-Net, each encoder block is directly connected to its corresponding decoder block through a simple skip connection, transferring low-level features for reconstruction. In contrast, U-Net++ adds multiple intermediate convolutional layers between encoder and decoder at various levels, creating a series of shorter, denser paths for information flow. These nested connections allow better feature refinement before upsampling, and the de-

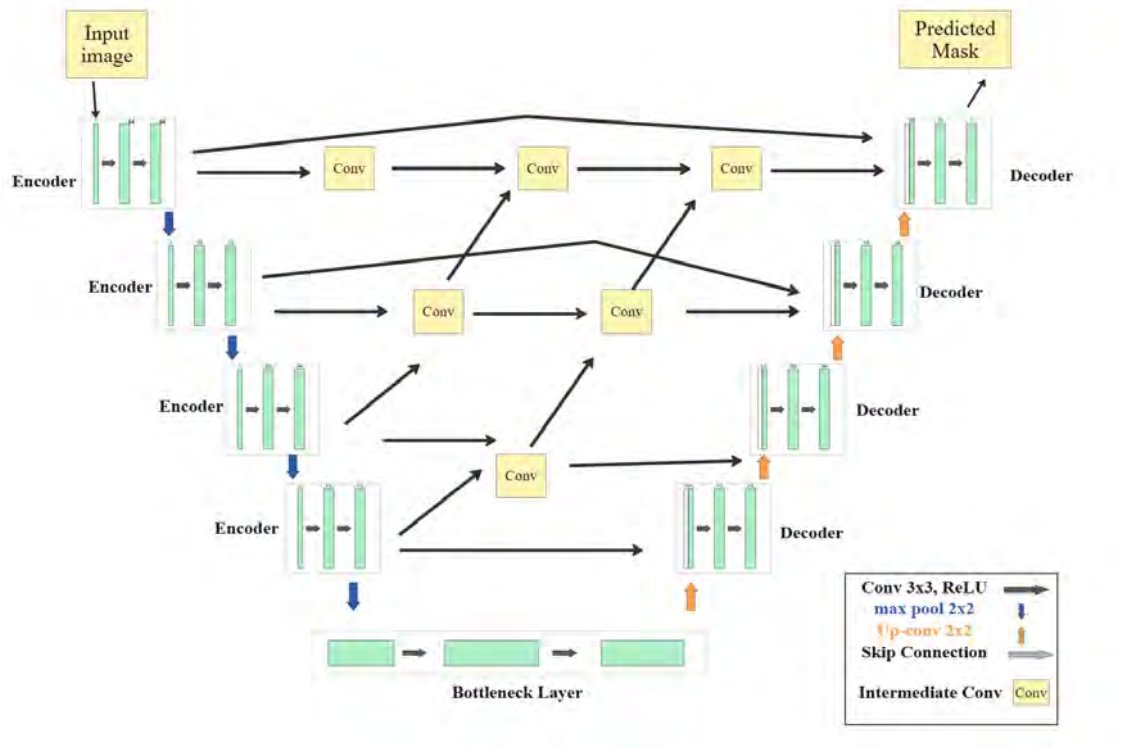

Figure 4.4: Unet++ model architecture

coder receives progressively enriched feature maps, making the model more expressive. While U-Net sends single-level skip connections, U-Net++ effectively reuses and merges multi-scale features, enabling more robust segmentation structures during training and inference.

4.3.3 ResUnet

The ResU-Net model integrates the U-Net architecture with residual connections, a concept borrowed from ResNet, to enhance gradient flow and enable deeper architectures without the common problem of vanishing gradients. In ResU-Net, each convolutional block within the encoder and decoder includes a residual connection, which means the input to a block is added to its output before activation. This helps the network learn identity mappings more effectively and allows for the training of deeper and more complex models. The overall structure still follows the encoder-decoder U-shaped design of U-Net, but with these added residual shortcuts, the model becomes more stable and efficient, especially on complex segmentation tasks like medical imaging.
In our implementation, ResU-Net operates on the Kvasir-SEG dataset using 3-channel RGB polyp images and corresponding binary masks. As the image passes through the encoder, each convolutional block not only processes features but also preserves the original input via residual links, which are then combined with the output. These preserved features help the model maintain low-level details that are essential for accurately segmenting the fine boundaries of polyps. In the decoder, upsampling layers reconstruct the segmentation map, and skip connections merge the encoder's features with the decoder's upsampled outputs. Unlike the plain U-Net

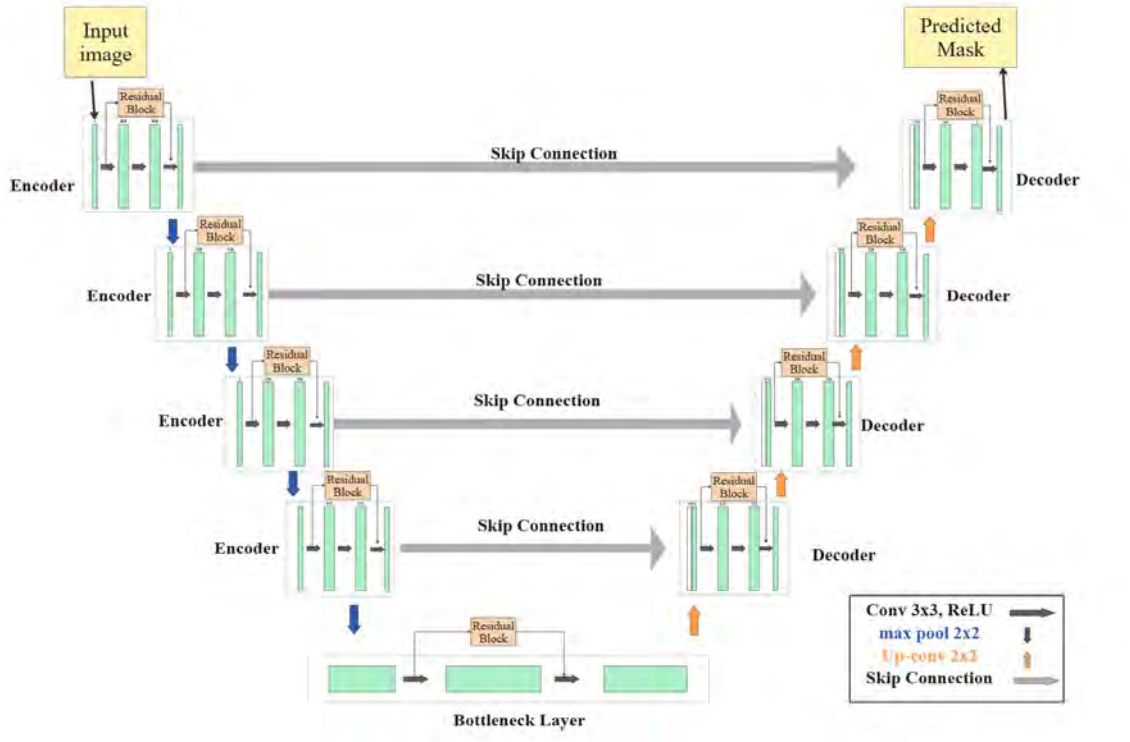

Figure 4.5: ResUnet model architecture

which might lose subtle spatial features during downsampling, ResU-Net's residual structure ensures that essential information is retained and reused effectively. This architectural enhancement allows the model to better capture complex textures and edges present in polyp images, making it particularly suitable for our dataset where fine-grained detail is crucial.

4.3.4 ResUnet++

The ResU-Net++ model builds upon the foundation of ResU-Net by incorporating additional architectural enhancements that aim to improve feature refinement and spatial information recovery. It combines three key concepts: residual connections, dense skip pathways, and attention mechanisms. Similar to ResU-Net, it maintains residual connections within its convolutional blocks to support better gradient flow and deeper network training. However, it goes a step further by integrating nested skip connections (inspired by U-Net++), which create intermediate dense pathways between encoder and decoder stages. These connections enhance multi-scale feature fusion, allowing the model to learn both local and global contextual information more effectively. Additionally, some variants may include attention gates that help the model focus on the most relevant regions during decoding.
In our implementation with the Kvasir-SEG dataset, ResU-Net++ processes RGB polyp images and binary segmentation masks through a more interconnected and enriched feature extraction pipeline compared to ResU-Net. While ResU-Net passes features through residual blocks and basic skip connections, ResU-Net++ introduces multiple intermediate connections between corresponding encoder and decoder layers, encouraging deeper supervision and better refinement at each stage. This results

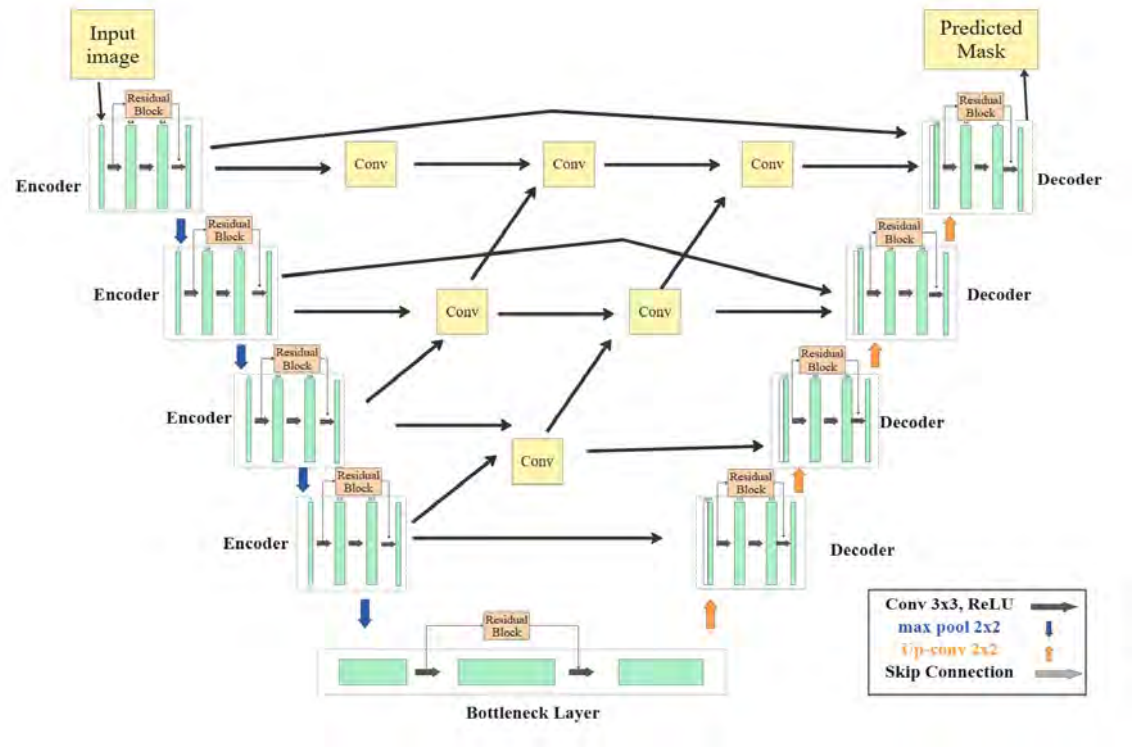

Figure 4.6: ResUnet++ model architecture

in more complex internal representations of the image, improving the network's ability to detect polyps with irregular shapes and faint boundaries. The nested pathways allow earlier and finer-level features to propagate through the network multiple times, which is particularly useful for our task where polyps vary significantly in size, shape, and visibility. Overall, the ResU-Net++ model's increased architectural depth and feature connectivity make it more capable of handling nuanced segmentation challenges, although it also requires more computational resources.

4.3.5 DeepLab V3

DeepLabV3 is a powerful semantic segmentation architecture that departs from the classic encoder-decoder structure of U-Net. Instead, it leverages dilated (atrous) convolutions to capture multi-scale contextual information without reducing spatial resolution. Its core innovation lies in the Atrous Spatial Pyramid Pooling (ASPP) module, which applies multiple parallel atrous convolutions with different dilation rates to extract features at various receptive fields. This allows the model to effectively capture both fine details and global context. DeepLabV3 does not explicitly decode the image back to its original size via a traditional decoder; rather, it uses bilinear upsampling after the ASPP module to restore the resolution. This design makes it computationally efficient while maintaining high accuracy for complex segmentation tasks.

In our polyp segmentation implementation, DeepLabV3 processes RGB input images and binary masks from the Kvasir-SEG dataset. The encoder backbone—ResNet-34 in our case—first extracts hierarchical features from the input image. These features are then passed to the ASPP module, which captures contextual information at

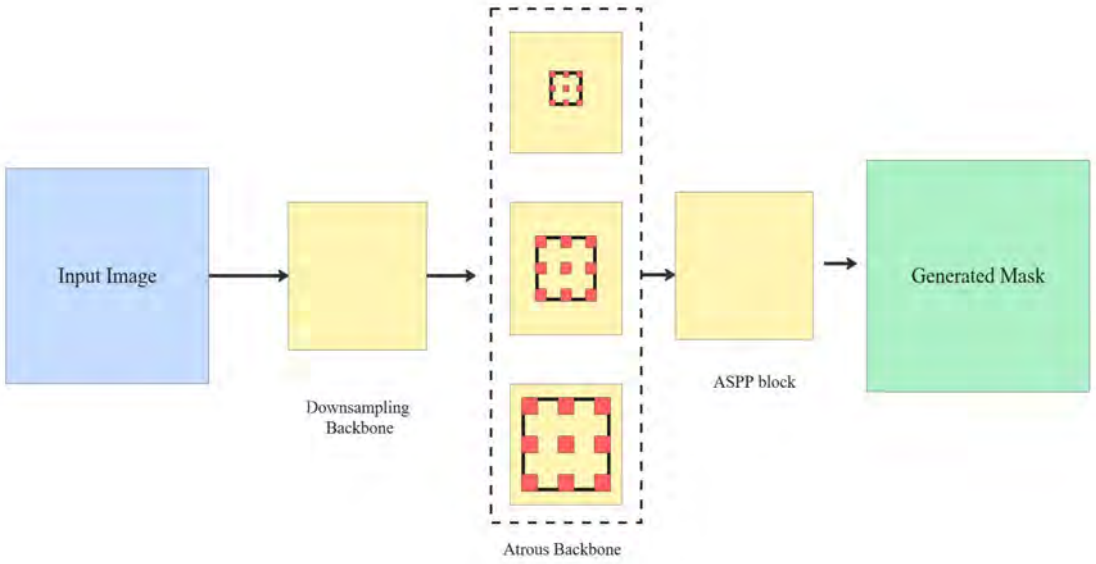

Figure 4.7: DeepLab V3 architecture

multiple scales. This is particularly beneficial for our dataset, as polyps vary widely in shape and size, and multi-scale features help the model detect both small, subtle lesions and larger, more prominent ones. After the ASPP module processes the feature maps, the output is upsampled to the original image size using bilinear interpolation to produce the final segmentation mask. DeepLabV3 is especially suitable for our task because it balances accuracy and efficiency, excels at preserving spatial detail, and is robust to scale variation, making it ideal for detecting polyps with diverse morphological characteristics.

4.3.6 FPN

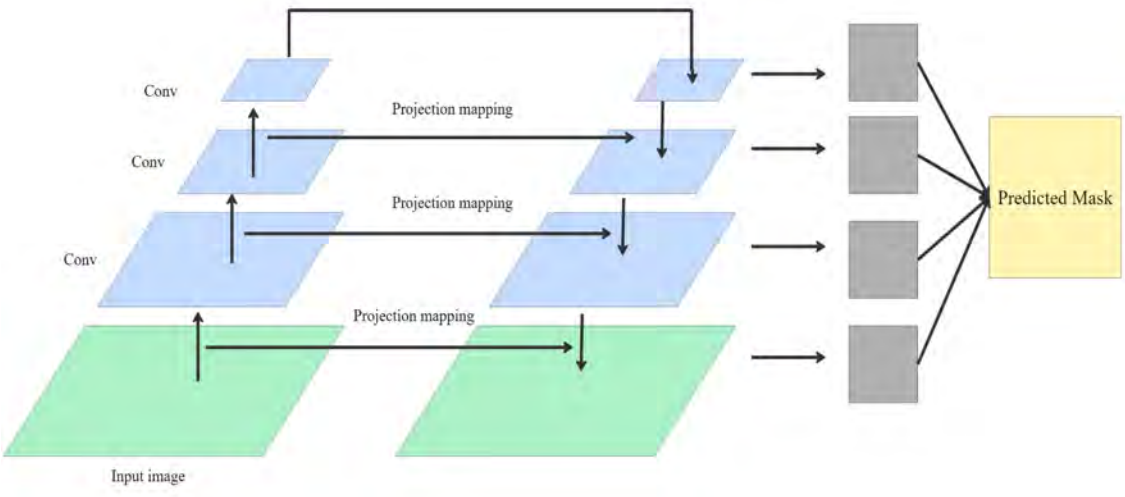

Figure 4.8: FPN model architecture

Feature Pyramid Network (FPN) is a segmentation architecture designed to extract and merge features at multiple scales in a top-down manner. It builds upon a backbone CNN (like ResNet-34 in our implementation) by creating a feature pyramid: instead of only using the final layer's output, it taps into multiple intermediate feature maps from different stages of the encoder. These multi-scale feature maps are then progressively upsampled and merged with higher-resolution features using lateral connections. This top-down path with lateral connections allows the network to combine semantically strong, low-resolution features with spatially rich, high-resolution features, enhancing the model's ability to localize and classify objects of varying sizes.
In our polyp segmentation setup, the RGB input images are passed through the ResNet-34 encoder, which extracts hierarchical features at different spatial resolutions. FPN then constructs a pyramid from these features by upsampling deeper layers and combining them with corresponding shallow layers via 1×1 convolutions. This fusion ensures that both fine-grained edge details and deep semantic context are retained. Once the pyramid is built, the fused features are processed to generate the segmentation mask, which is upsampled to the original resolution. This architecture proves especially useful in our dataset where polyps appear at various scales and with blurred boundaries—FPN's multi-scale fusion helps detect small lesions more accurately and maintain edge clarity, contributing to improved segmentation quality while keeping the model relatively lightweight.

4.3.7 Swin-Unet V2

The segmentation model employed in this study leverages a hierarchical feature extraction approach based on the Swin Transformer V2 architecture, which has demonstrated strong performance in various vision tasks. Unlike traditional convolutional neural networks like U-Net, which rely on convolutional layers and skip connections to capture multiscale features, this model uses a transformer-based backbone to encode global and local context through self-attention mechanisms. The Swin Transformer processes input images in a patch-wise manner across multiple encoder layers, producing rich, multi-scale feature representations. These are then progressively upsampled and fused through a custom decoder composed of specialized linear upsampling modules, designed to restore spatial resolution and combine semantic details across scales. The final output is a high-resolution segmentation mask generated via convolution and sigmoid activation, enabling pixel-wise classification for accurate polyp detection.
In our implementation on the Kvasir-SEG dataset, each input image and its corresponding mask are resized and transformed into tensors compatible with the Swin Transformer backbone. The model extracts features at five distinct levels corresponding to different encoder stages, capturing hierarchical representations of polyp morphology with improved contextual awareness compared to convolution-based models. These multi-level features are fed into the decoder network, which sequentially upsamples and merges them to reconstruct spatial details essential for precise segmentation. Compared to U-Net's convolutional encoder-decoder design with symmetric skip connections, the transformer-based approach excels at modeling long-range dependencies and global context, which can improve segmentation accu-

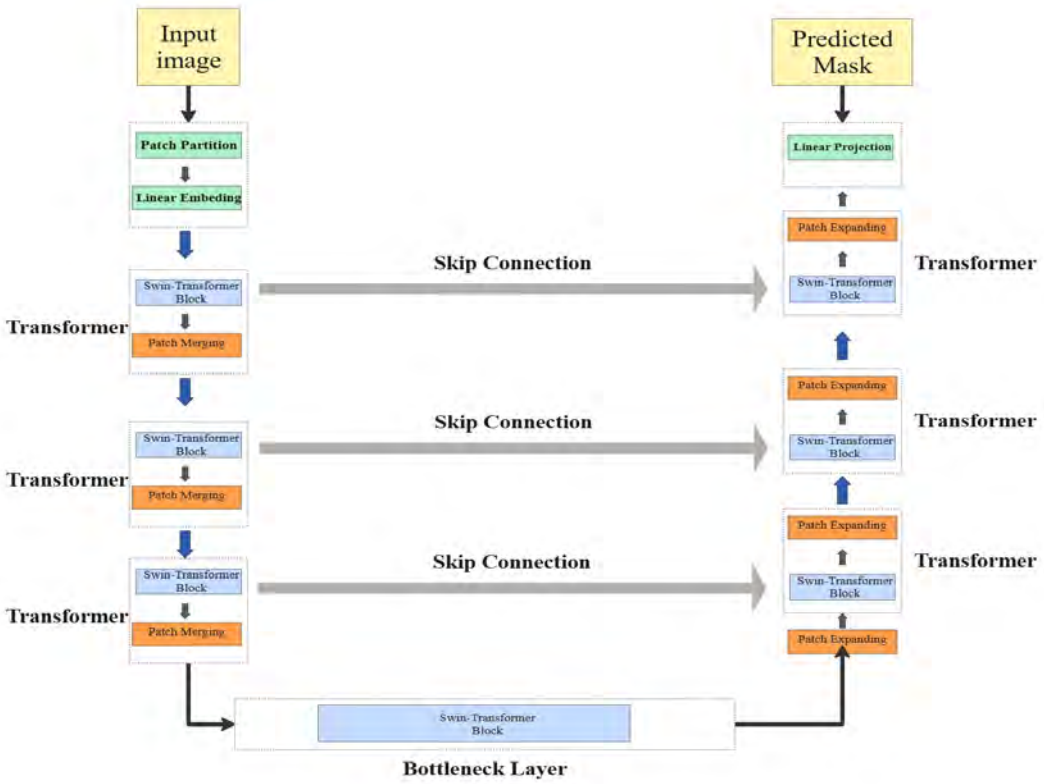

Figure 4.9: Swin-Unet V2 model architecture

racy on complex polyp structures. Training uses binary cross-entropy loss optimized via Adam, enabling the network to effectively learn mappings from transformer-encoded features to segmentation masks that closely match ground truth annotations.

Chapter 5

Results and Analysis

5.1 Performance Evaluation Metrics

To evaluate the performances of the models, we have recorded multiple performance metrics of these models. After training the models on training datasets, we tested the models' performances by running them on test datasets and recording these values.

5.1.1 Average Precision and Recall

Precision is a metric that measures the accuracy of positive predictions. It is the ratio of correctly predicted positive observations to the total predicted positives which is calculated by using equation (1). on the other hand, Recall (Sensitivity) is the ability of the model to correctly identify all relevant positive instances. It is the ratio of correctly predicted positive observations to the total actual positives and is calculated with the help of equation (2). The parameter TP is True Positives, FP is False Positive and FN is False Negatives in these equations.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

5.1.2 Average Specificity

Specificity is the proportion of actual negatives that are correctly identified as such. It answers the question "Out of all the negative instances, how many did the model correctly identify?" We used equation (3) to calculate the values. Here, TN is True Negative and FP is false positive.

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{3}$$

5.1.3 Dice Coefficient

The Dice Coefficient (also known as the Dice Similarity Coefficient, DSC) is a measure of overlap between the predicted and ground truth segmentation masks. By using equation (5) we can calculate the coefficient. A higher Dice coefficient means

a better match between the two.

$$\text{Dice Coefficient} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{5}$$

5.1.4 Intersection over Union (IoU)

IoU basically gives the overlap between the predicted segmentation mask and the ground truth mask. The equation (6) gives a way to calculate the IoU value. This is an important metric when considering the segmentation performance because it penalizes both false positives and false negatives.

$$\text{IoU} = \frac{TP}{TP + FP + FN} \tag{6}$$

5.1.5 F1 Score

The F1-Score is the harmonic mean of precision and recall, hence a balanced metric that considers both false positives and false negatives. It is useful, especially in cases where the class distribution is imbalanced. We compute the F1 score using equation (7) and the values of the parameters; Precision by equation (1) and recall is calculated using equation (2).

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

5.2 Analysis of Default Model Performances

Metric	UNet	UNet++	ResUNet	ResUNet++	DeepLabV3	FPN	SwinUnet V2
Precision	0.8848	**0.9431**	0.9330	0.9271	0.9484	0.9258	0.8738
Recall	0.8651	**0.8880**	0.8703	0.8425	0.8362	0.8795	0.8403
Specificity	0.9790	0.9906	0.9898	0.9877	**0.9910**	0.9880	0.9537
Dice	0.8733	**0.9140**	0.8979	0.8817	0.8871	0.9010	0.8494
IoU	0.7797	**0.8433**	0.8167	0.7905	0.7994	0.8209	0.7382
F1 Score	0.8733	**0.9140**	0.8979	0.8817	0.8871	0.9010	0.8375
Training Time (s)	421.06	497.96	426.85	596.57	479.77	422.77	623.23

Table 5.1: Comparison of different segmentation models on various performance metrics.

5.3 Comparative Analysis of Model Performance

The performance evaluation of seven deep learning-based polyp segmentation models—**UNet**, **UNet++**, **ResUNet**, **ResUNet++**, **DeepLabV3**, **FPN**, and **SwinUnet** V2 revealed distinct strengths and trade-offs across multiple metrics. These include Dice coefficient, IoU, Precision, Recall, Specificity, F1-score, and Training Time. Each metric plays a crucial role in the context of medical image segmentation, especially for clinical applications such as early colorectal cancer detection via colonoscopy.

- **UNet++** emerged as the top performer, achieving the highest Dice score (0.9140), IoU (0.8433), and Recall (0.8880). Its nested and dense skip connections enable superior multiscale feature fusion and edge sensitivity, making it ideal for segmenting polyps with complex boundaries.
- **DeepLabV3** reported the highest Precision (0.9484) and Specificity (0.9910), due to its use of Atrous Spatial Pyramid Pooling (ASPP). These features are effective in suppressing false positives but resulted in a lower Recall (0.8362), potentially missing subtle polyps.
- **ResUNet** delivered balanced performance with Dice (0.8979) and IoU (0.8167), benefiting from residual blocks that stabilize training and enhance feature reuse. It offers a well-rounded model suitable for general deployments.
- **FPN** showed unexpectedly strong results with Dice (0.9010) and Recall (0.8795). Its top-down, lateral fusion strategy supports effective segmentation of polyps at different scales, and it trained efficiently with a time of 422.77 seconds.
- **ResUNet++** performed moderately well across all metrics, though its complex structure, including ASPP and SE modules, did not yield a proportional increase in segmentation quality under default settings.
- **UNet**, while the oldest model, achieved a respectable Dice (0.8733) and the fastest training time (421.06s). It remains a viable choice in low-resource or real-time environments, albeit less capable of fine boundary detection.

- **Swin-Unet V2**, the transformer-based architecture, under-performed (Dice = 0.8494; IoU = 0.7382) and had the second-highest training time. This model is likely under-optimized and requires significant fine-tuning and larger datasets to reach its potential.

In summary, **UNet++** provides the best balance of performance metrics, making it the optimal choice for polyp segmentation tasks in clinical practice. However, models like **DeepLabV3** and **FPN** also demonstrate valuable capabilities in precision-focused and resource-sensitive deployments, respectively. No model was universally superior across all metrics, highlighting the importance of metric prioritization based on application context. The transformer-based model, **Swin-Unet v2** under-performed in almost all parameters and also took way more train time. This is due to the over-complexity of transformer models which also requires very large train datasets.
Therefore we now proceed to fine tune the rest of the models and evaluate their performance changes.

5.4 Final Design Adjustment

5.4.1 Fine-Tuning Strategy

To enhance the performance of all segmentation models explored in this study: U-Net, U-Net++, ResUNet, ResUNet++, DeepLabV3, FPN, and the Transformer-based architecture—a robust fine-tuning strategy was implemented. This consisted of improvements in three critical areas: the encoder backbone, the optimizer, and the learning rate scheduling mechanism.

Encoder Improvement

All models were upgraded to use **EfficientNet-B4** as the encoder backbone instead of conventional architectures like ResNet-34. EfficientNet-B4 leverages compound scaling to uniformly scale depth, width, and resolution, offering a better accuracy-efficiency trade-off. By using ImageNet-pretrained weights, the encoder was able to extract rich and abstract features from medical images, which is crucial for precise segmentation tasks involving subtle anatomical structures.

Advanced Optimization

The standard Adam optimizer was replaced with **AdamW**, which decouples weight decay from the optimization step. This modification improves generalization and regularization, particularly beneficial when working with limited or noisy data. A weight decay value of 1×10^{-4} was used to penalize large weights and promote sparsity in model parameters.

Learning Rate Scheduling

A **Cosine Annealing Learning Rate (CosineAnnealingLR)** scheduler was employed to allow gradual and smooth convergence. The scheduler reduces the learning

rate following a cosine decay curve, preventing abrupt changes that could destabilize training. The parameters were set as $T_{max} = 10$ and $\eta_{min} = 10^{-6}$ to balance learning speed and convergence accuracy.
This comprehensive fine-tuning strategy contributed significantly to improved training stability, convergence speed, and segmentation accuracy across all models.

Metric	**U-Net**	**U-Net++**	**ResUNet**	**ResUNet++**	**DeepLabV3**	**FPN**
Precision	0.9242	0.9558	0.9405	0.9339	0.9060	0.9333
Recall	0.9063	0.8885	0.8810	0.8292	0.9225	0.9047
Specificity	0.9857	0.9922	0.9891	0.9890	0.9807	0.9871
Dice	0.9139	0.9200	0.9083	0.8767	0.9131	0.9178
IoU	0.8426	0.8535	0.8337	0.7833	0.8417	0.8491
F1-Score	0.9722	0.9744	0.9711	0.9636	0.9709	0.9733

Table 5.2: Performance Metrics Afer Fine-tune and Loss Function Implementation

The table 5.2 give the performace metrics values generated by the models after **fine tuning** and implementing novel **boundary-aware dual loss** learning rate. The table is missing **Swin-Unet V2** model values as we did not implement fine tuning as it under-performed in our defaut analysis as mentioned in previous section.

5.4.2 Boundary-Aware Dual Loss Function

In addition to architectural and training improvements, a custom loss function termed **Boundary-Aware Dual Loss** was introduced. This function is a combination of **Dice Loss** and a novel **Boundary Focal Loss**, designed to address both region-level and boundary-level segmentation challenges.
The default model uses Binary Cross Entropy Loss function which is usually used for binary pixel classification task. However, our goal is to segment polyp affected area from the input images which comes in various shapes and sizes. To guide the model into segmenting polyp properly, our proposed **Boundary-Aware Dual Loss Function** combines them to capture better segmentation.

Dice Loss

Dice Loss directly maximizes the overlap between predicted and ground truth masks, making it highly effective for handling class imbalance. It ensures that the model prioritizes accurate coverage of relevant regions (e.g., polyps or lesions), which is critical in medical image segmentation.

Boundary Focal Loss

To enhance edge precision, a Boundary Focal Loss was implemented by applying **Canny edge detection** to the ground truth masks to extract boundary pixels. These boundary regions are then weighted using a focal variant of Binary Cross-Entropy (BCE) loss. This approach emphasizes difficult-to-classify edge pixels, encouraging the model to produce smoother and more anatomically faithful boundaries.

Combined Loss Formulation

The final loss is a weighted sum of the two components:

$$\mathcal{L}_{\text{total}} = \alpha \cdot \mathcal{L}_{\text{Dice}} + \beta \cdot \mathcal{L}_{\text{BoundaryFocal}}$$

Empirically, α and β were set to 0.7 and 0.3 respectively, to favor region overlap while still enforcing boundary sharpness. We have tried a few combination of α and β and finalized this values for optimal performance. This combination helped produce segmentation masks that are both globally accurate and locally precise, offering a balanced solution for complex medical image segmentation tasks.

5.5 Comparisons and Relationships

Model	**Dice Δ**	**IoU Δ**	**Precision Δ**	**Recall Δ**
U-Net	+0.0406	+0.0629	+0.0394	+0.0412
U-Net++	+0.0060	+0.0102	+0.0127	unchanged
ResUNet	+0.0104	+0.0170	+0.0075	+0.0107
DeepLabV3	+0.0260	+0.0423	-0.0424	+0.0863
FPN	+0.0168	+0.0282	+0.0075	+0.0252
ResUNet++	-0.0050	-0.0072	unchanged	-0.0133

Table 5.3: Summary of Performance Improvements After Applying Custom Loss and Fine-Tuning

The application of a custom loss function combined with fine-tuning strategies significantly enhanced the performance of various segmentation models for polyp detection. The custom loss, likely a hybrid of Dice Loss and Binary Cross-Entropy or a similar formulation, addressed common challenges in medical image segmentation such as class imbalance and boundary ambiguity. By integrating region-based overlap metrics with pixel-wise accuracy penalties, this loss encouraged models to better capture the fine details and edges of polyps, improving overall segmentation quality.

- Among the evaluated models, **U-Net++** continued to demonstrate the best overall performance. Its Dice coefficient improved modestly from 0.9140 to 0.9200, and Intersection over Union (IoU) rose from 0.8433 to 0.8535. Notably, its Precision and Specificity increased to 0.9558 and 0.9922 respectively, highlighting its superior ability to discriminate polyp boundaries and reduce false positives. The Recall metric remained stable, indicating that the model did not compromise true positive detection in pursuit of higher precision. These improvements can be attributed to the custom loss's emphasis on boundary refinement and penalization of over-segmentation, which synergized well with U-Net++'s nested architecture.
- The **Feature Pyramid Network (FPN)** exhibited a balanced improvement across all metrics, with Dice increasing to 0.9178 and IoU to 0.8491. Precision and Recall both saw meaningful increases, suggesting that FPN's multi-scale

feature aggregation benefited from the loss function's pixel and region-focused penalties. This synergy enabled more robust segmentation of polyps at varying scales and complexities.

- The **DeepLabV3** model showed a remarkable increase in Recall, rising sharply from 0.8362 to 0.9225, the highest among all tested models. While Dice improved moderately to 0.9131, Precision saw a slight decrease. This indicates that the custom loss function encouraged the model to capture more potential polyp pixels, effectively reducing missed detections at the cost of a small increase in false positives—a trade-off often desirable in clinical applications where missing lesions is more critical than over-segmentation.

- The baseline **U-Net** model also benefited substantially from the new loss and fine-tuning, with Dice improving from 0.8733 to 0.9139 and IoU from 0.7797 to 0.8426. Precision rose notably, reflecting better foreground-background discrimination despite the simpler architecture, showing that the hybrid loss helped compensate for the lack of complex refinement stages.

- **ResUNet** experienced a steady improvement across metrics, demonstrating robust segmentation without overfitting. Dice and IoU increased modestly, supported by gains in both Precision and Recall. The residual connections in ResUNet, combined with the custom loss's pixel-level guidance, enhanced the model's ability to segment polyps accurately and consistently.

- However, the **ResUNet++** model's performance declined slightly after applying the custom loss. Dice dropped to 0.8767 and IoU to 0.7833, with Recall falling to the lowest among all models at 0.8292, although Precision remained relatively high. This suggests that the model may have suffered from an architectural mismatch with the loss function, where the complex decoder components (such as ASPP and SE blocks) conflicted with the new loss dynamics, resulting in under-segmentation or instability.

In summary, the custom hybrid loss function, by addressing class imbalance and emphasizing boundary accuracy through combined Dice and BCE terms, facilitated more accurate and clinically useful polyp segmentation across most architectures. It improved the critical metrics of Recall and Dice coefficient, which are vital to minimizing missed lesions. U-Net++ maintained its lead, while FPN and DeepLabV3 also showed meaningful advancements in balancing precision and recall. The only notable exception was ResUNet++, where architectural complexity may have impeded synergy with the loss.

5.6 Summary of the results

We have examined both the base and fine-tuned versions of prominent segmentation models, namely UNet, UNet++, ResUNet, ResUNet++, DeepLabV3, Feature Pyramid Network (FPN), and SwinUNet. The evaluation is based on standard computer vision metrics including Dice coefficient, Intersection over Union (IoU), F1-score, Precision, Recall (Sensitivity), and Specificity. Upon analyzing the results of both sets (base and fine-tuned), we identify the top five best-performing models

overall based on a balance of segmentation accuracy, recall, and precision which are UNet++ (Fine-Tuned), FPN (Fine-Tuned), UNet (Fine-Tuned), DeepLabV3 (Fine-Tuned), ResUNet (Fine-Tuned) UNet++ (fine-tuned) exhibits the highest Dice coefficient (0.9200) and IoU (0.8535) among all models, making it the most accurate in terms of overlap with ground truth. It also achieves the highest F1-score (0.9744), indicating strong harmonic balance between precision (0.9558) and recall (0.8885). Moreover, it demonstrates excellent specificity (0.9922), ensuring minimal false positives. These combined make it the strongest candidate for clinical-grade polyp segmentation. The fine-tuned FPN model follows closely behind with a Dice coefficient of 0.9178, IoU of 0.8491, and F1-score of 0.9733. It also shows high precision (0.9333) and recall (0.9047), meaning it is both sensitive and accurate. Its specificity of 0.9871 indicates strong ability to distinguish background from polyps, making it ideal for high-recall applications like early diagnosis systems. UNet (fine-tuned) ranks third with a Dice of 0.9139 and a notable F1-score of 0.9722. It has the highest recall (0.9063) after DeepLabV3, along with excellent specificity (0.9857) and a solid IoU of 0.8426. This makes it a great candidate when both false positives and false negatives must be minimized. DeepLabV3's fine-tuned version stands out for achieving the highest recall (0.9225), suggesting that it detects most polyp regions without missing subtle ones. Although its precision is slightly lower (0.9060), the model maintains a strong Dice (0.9131) and IoU (0.8417). It is especially valuable in medical scenarios where missing any part of the polyp is critical. ResUNet's fine-tuned variant achieves a Dice of 0.9083, IoU of 0.8337, and F1-score of 0.9711, showing strong performance. With a precision of 0.9405 and specificity of 0.9891, it is highly precise and conservative in segmentation, which can be beneficial when minimizing over-segmentation is essential.

The base models showed competitive results but were consistently outperformed by their fine-tuned counterparts. Notably, UNet++ (Base) was the strongest among base models with a Dice of 0.9140 and IoU of 0.8433, outperforming many others even before tuning. DeepLabV3 (Base) had the highest precision (0.9484) among all base models, though its recall was comparatively lower (0.8362). SwinUNet (Base) lagged behind in most metrics with the lowest Dice (0.8494) and IoU (0.7382), despite having a long training time. Fine-tuning significantly improves model performance, particularly for UNet++, FPN, and UNet, which top the rankings across all evaluated metrics. Based on a holistic evaluation, UNet++ (Fine-Tuned) stands out as the best model for polyp segmentation due to its superior balance of precision, recall, and segmentation overlap. For real-world applications requiring robust and accurate segmentation, the fine-tuned versions of UNet++, FPN, UNet, DeepLabV3, and ResUNet offer the most promise.

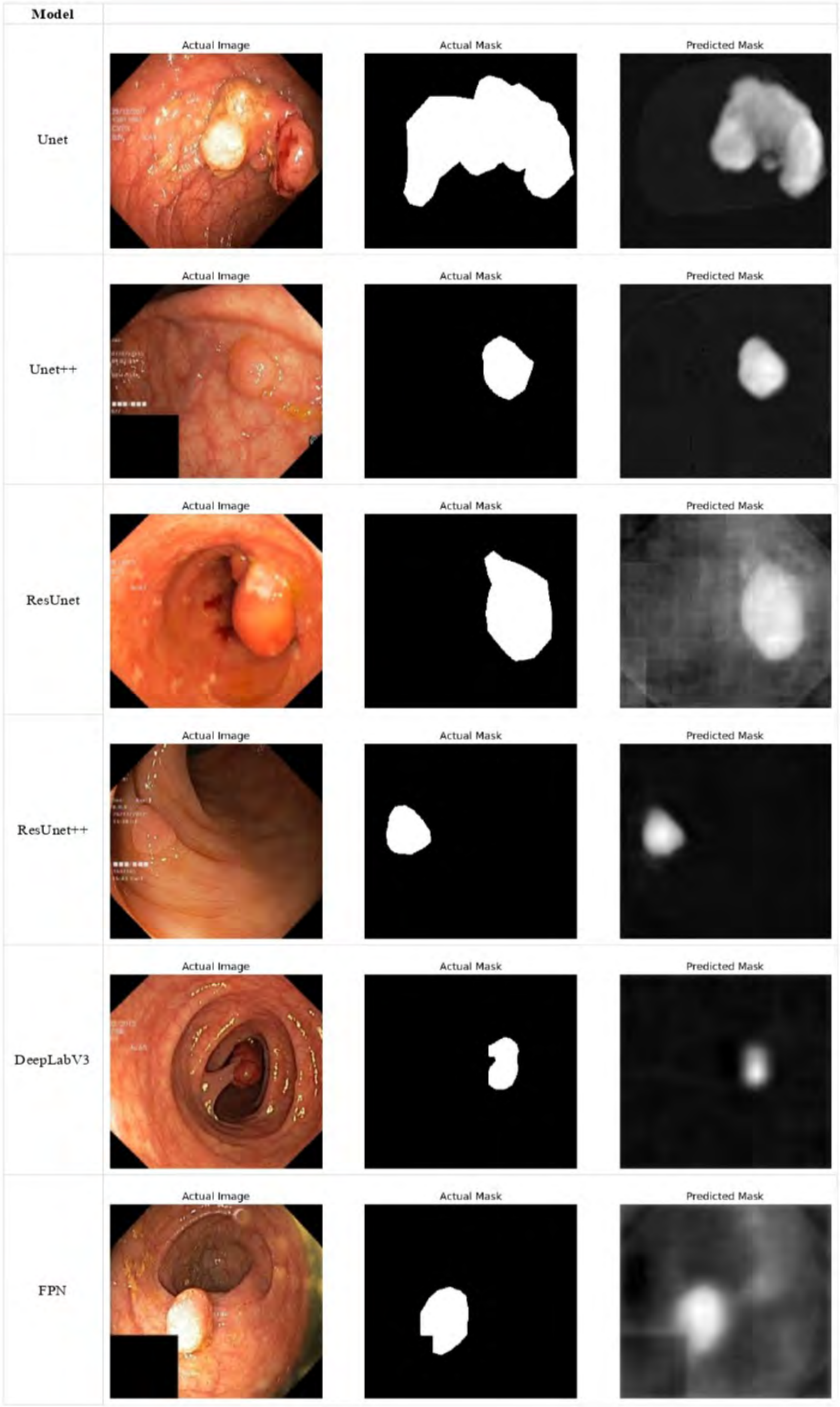

Figure 5.1: Actual mask vs predicted mask by the models

Chapter 6

Conclusion

6.1 Economic Analysis

- **Reduction in Healthcare Costs:** Early and accurate detection of colorectal polyps through automated deep learning models can significantly reduce long-term healthcare expenses. By identifying precancerous polyps before they develop into colorectal cancer, costly treatments such as chemotherapy, surgeries, or prolonged hospital stays can be avoided.
- **Improved Diagnostic Efficiency:** Implementing automated polyp segmentation in hospitals and clinics leads to faster and more consistent diagnostics, reducing the workload on gastroenterologists. This efficiency translates into reduced diagnostic labor costs and enables clinicians to attend to more patients, increasing throughput without increasing staff size.
- **Minimizing Diagnostic Errors:** Human oversight in polyp detection can lead to missed cases, which result in advanced-stage treatments later on. Reducing false negatives with AI-enhanced diagnostics prevents these downstream costs, thereby improving cost-effectiveness in medical imaging.
- **Scalability for Underserved Regions:** Low-resource or rural healthcare facilities often lack access to experienced radiologists or specialists. Deploying these models in such areas can democratize access to quality diagnostics without the need for expensive human capital, contributing to equitable healthcare access and boosting labor productivity in the healthcare sector.
- **Boosting Medical AI Sector:** As models like U-Net++ become clinically validated, there is increased demand for AI development, data annotation services, and specialized hardware—supporting job creation and investment in the growing medical AI economy.
- **Insurance and Public Health Savings:** Earlier detection rates reduce the average claim sizes for insurance companies and lessen the burden on public healthcare systems, allowing better allocation of national health budgets toward preventive care rather than emergency interventions.

6.2 Ethical Issues

- **Data Privacy and Patient Consent:** Medical imaging data used to train deep learning models often originates from real patients. If proper anonymiza-

tion and informed consent protocols are not followed, there is a risk of violating patient privacy and ethical standards in data handling.

- **Bias and Fairness in Models:** Deep learning models trained on datasets from specific populations (e.g., Kvasir-SEG based on Norwegian patients) may underperform when applied to patients from other ethnic or demographic backgrounds, leading to unequal diagnostic outcomes and potential health disparities.
- **Transparency and Explainability:** These AI models often function as "black boxes," making it difficult for clinicians to understand how a decision was reached. This lack of transparency can lead to ethical dilemmas in clinical accountability, especially when an AI-generated result is used to justify invasive procedures.
- **Overreliance on Automation:** Dependence on automated systems could lead to reduced vigilance or skill erosion among medical professionals. If clinicians begin to trust AI outputs blindly without human oversight, it could increase the risk of misdiagnosis or missed detection in edge cases.
- **Regulatory and Liability Challenges:** In cases where AI systems make incorrect predictions that harm patients, it becomes ethically complex to determine liability—should responsibility fall on the developers, the healthcare provider, or the institution?
- **Access and Equity:** Advanced AI tools may only be available in well-funded hospitals or regions, widening the gap in healthcare access between rich and poor communities or countries. Ethical concerns arise when life-saving technologies are not equitably distributed.
- **Use Beyond Intended Purpose:** Once trained, AI models could be adapted for non-medical or unauthorized uses without consent or oversight, raising questions about the ethical scope of deployment and potential misuse.

6.3 Conclusion

In this study, we explored and evaluated the efficacy of multiple deep learning architectures for the segmentation and identification of polyps from endoscopic images—an essential step toward the early detection of colorectal cancer. Using the Kvasir-SEG dataset, we implemented and benchmarked seven models: U-Net, U-Net++, ResUNet, ResUNet++, DeepLabV3, Feature Pyramid Network (FPN), and Swin-Unet V2, under a standardized experimental framework with consistent preprocessing, augmentation, and evaluation metrics.

Among the models tested, U-Net++ emerged as the top performer with a Dice Coefficient of 0.9140 and an IoU of 0.8433, demonstrating its superior ability to localize complex polyp boundaries. Transformer-based Swin-Unet V2 also showcased promising results in capturing long-range dependencies, although it came with increased computational cost. Our comparative analysis highlighted the importance of architectural innovations such as residual connections, multi-scale feature extraction, and attention mechanisms in improving segmentation accuracy.

Additionally, we observed that applying a customized boundary-aware loss function further refined segmentation precision, particularly at polyp edges—an area where many models struggle due to the variability in texture, shape, and contrast of the target regions. Despite these advances, challenges remain in achieving real-time deployment and ensuring robust performance across diverse datasets, especially in resource-limited clinical settings.

This research not only validates the effectiveness of modern segmentation networks for polyp detection but also underscores the need for lightweight, generalizable, and interpretable models. Future work will aim to integrate more diverse datasets, refine hybrid architectures (e.g., combining transformers with CNNs), and explore ensemble approaches to enhance diagnostic performance. Ultimately, this study serves as a foundational step toward building reliable, AI-driven diagnostic support tools that can be confidently deployed in routine colonoscopy screenings.

Bibliography

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., Cham: Springer International Publishing, 2015, pp. 234–241, ISBN: 978-3-319-24574-4.

[2] P. Brandao, E. Mazomenos, G. Ciuti, *et al.*, "Fully convolutional neural networks for polyp segmentation in colonoscopy," in *Medical Imaging 2017: Computer-Aided Diagnosis*, S. G. A. III and N. A. Petrick, Eds., International Society for Optics and Photonics, vol. 10134, SPIE, 2017, 101340F. DOI: 10.1117/12.2254361. [Online]. Available: https://doi.org/10.1117/12.2254361.

[3] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, *Rethinking atrous convolution for semantic image segmentation*, 2017. arXiv: 1706.05587 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1706.05587.

[4] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-denseunet: Hybrid densely connected unet for liver and tumor segmentation from ct volumes," *IEEE Transactions on Medical Imaging*, vol. 37, no. 12, pp. 2663–2674, 2018. DOI: 10.1109/TMI.2018.2845918.

[5] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, *Unet++: A nested u-net architecture for medical image segmentation*, 2018. arXiv: 1807.10165 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1807.10165.

[6] D. Jha, P. H. Smedsrud, M. A. Riegler, *et al.*, "Resunet++: An advanced architecture for medical image segmentation," in *2019 IEEE International Symposium on Multimedia (ISM)*, 2019, pp. 225–2255. DOI: 10.1109/ISM46123.2019.00049.

[7] Y. Weng, T. Zhou, Y. Li, and X. Qiu, "Nas-unet: Neural architecture search for medical image segmentation," *IEEE Access*, vol. 7, pp. 44 247–44 257, 2019. DOI: 10.1109/ACCESS.2019.2908991.

[8] D.-P. Fan, G.-P. Ji, T. Zhou, *et al.*, "Pranet: Parallel reverse attention network for polyp segmentation," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, A. L. Martel, P. Abolmaesumi, D. Stoyanov, *et al.*, Eds., Cham: Springer International Publishing, 2020, pp. 263–273, ISBN: 978-3-030-59725-2.

[9] S. Hao, Y. Zhou, and Y. Guo, "A brief survey on semantic segmentation with deep learning," *Neurocomputing*, vol. 406, pp. 302–321, 2020, ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2019.11.118. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231220305476.

[10] H. Huang, L. Lin, R. Tong, *et al.*, "Unet 3+: A full-scale connected unet for medical image segmentation," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 1055–1059. DOI: 10.1109/ICASSP40776.2020.9053405.

[11] L. F. Sánchez-Peralta, L. Bote-Curiel, A. Picón, F. M. Sánchez-Margallo, and J. B. Pagador, "Deep learning to find colorectal polyps in colonoscopy: A systematic literature review," *Artificial Intelligence in Medicine*, vol. 108, p. 101 923, 2020, ISSN: 0933-3657. DOI: https://doi.org/10.1016/j.artmed.2020.101923. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0933365719307493.

[12] S. Ali, M. Dmitrieva, N. Ghatwary, *et al.*, "Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy," *Medical Image Analysis*, vol. 70, p. 102 002, 2021, ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2021.102002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841521000487.

[13] D. Jha, S. Ali, N. K. Tomar, *et al.*, "Real-time polyp detection, localization and segmentation in colonoscopy using deep learning," *IEEE Access*, vol. 9, pp. 40 496–40 510, 2021. DOI: 10.1109/ACCESS.2021.3063716.

[14] S. Safarov and T. K. Whangbo, "A-denseunet: Adaptive densely connected unet for polyp segmentation in colonoscopy images with atrous convolution," *Sensors*, vol. 21, no. 4, 2021, ISSN: 1424-8220. DOI: 10.3390/s21041441. [Online]. Available: https://www.mdpi.com/1424-8220/21/4/1441.

[15] R. Wang, T. Lei, R. Cui, B. Zhang, H. Meng, and A. K. Nandi, "Medical image segmentation using deep learning: A survey," *IET Image Processing*, vol. 16, no. 5, pp. 1243–1267, 2022. DOI: https://doi.org/10.1049/ipr2.12419. [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/ipr2.12419.

[16] H. Cao, Y. Wang, J. Chen, *et al.*, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *Computer Vision – ECCV 2022 Workshops*, L. Karlinsky, T. Michaeli, and K. Nishino, Eds., Cham: Springer Nature Switzerland, 2023, pp. 205–218, ISBN: 978-3-031-25066-8.

[17] World Health Organization, *Cancer*, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer.