Arab American University
Faculty of Graduate Studies
Department of Natural, Engineering, and Technology Sciences
Master Program in Data Science and Business Analytics

Polyp Detection Using Vision Transformers

Asma Alawneh
201912869

Supervision Committee:
Dr. Mohammad Yaqub
Dr. Majdi Owda
Prof. Hazem Hiary

This Thesis Was Submitted in Partial Fulfilment of the Requirements for the Master Degree in Data Science and Business Analytics

Palestine, April/2025

Arab American University
Faculty of Graduate Studies
Department of Natural, Engineering, and Technology Sciences
Master Program in Data Science and Business Analytics

Thesis Approval
Polyp Detection Using Vision Transformers

Asma Alawneh
201912869

This thesis was defended successfully on 06/04/2025 and approved by:

Thesis Committee Members:

Name	Title	Signature
1. Dr. Mohammad Yaqub	Main Supervisor	……..
2. Dr. Majdi Owda	Member of Supervision Committee	……..
3. Prof. Hazem Hiary	Member of Supervision Committee	……..

Palestine, April/2025

Declaration

I declare that, except where explicit reference is made to the contribution of others, this thesis is substantially my own work and has not been submitted for any other degree at the Arab American University or any other institution.

Student Name: Asma Ahmad Amin Alawneh

Student ID: 201912869

Signature: Asma Alawneh

Date of Submitting the Final Version of the Thesis: 14/08/2025

Acknowledgments

I would like to express my gratitude to my supervisor, Dr. Mohammad Yaqub, for the continuous support and valuable feedback throughout the process of thesis completion. I am also grateful to the Arab American University - Palestine, for providing me with the opportunity to work on this thesis. Finally, I extend my deepest appreciation to my family for their continuous love and support.

Polyp Detection Using Vision Transformers

Asma Ahmad Ameen Alawneh

Dr. Mohammad Yaqub

Dr. Majdi Owda

Prof. Hazem Hiary

Abstract

Colorectal cancer is the second leading cause of cancer deaths worldwide, and early detection and removal of polyps is crucial in lowering mortality rates. Unfortunately, traditional manual analysis of colonoscopy images is time-consuming, subjective, and can result in miss rates between 22-28%. This thesis investigates Vision Transformers as an innovative technique for semantic segmentation of polyps in colonoscopy images while comparing their performance against traditional Convolutional Neural Networks (CNNs).

The implementation pipeline involved data preparation steps such as splitting and online augmentation, followed by training and tuning of the segmentation models. Five widely used datasets (Kvasir, CVC-ClinicDB, CVC-ColonDB, EndoScene, and ETIS) were utilized for training and evaluation purposes. CNN-based models (DeepLabV3 with ResNet backbone, U-Net, and their ensembles) were implemented as a baseline, and the state-of-the-art Vision Transformer models (SegFormer, UperNet with Swin Transformer backbone and their ensembles) were implemented. All implemented models were tested against both seen and unseen datasets to gain insight into their generalization ability for real-world clinical applications.

During validation, the Ensemble_CNNs model achieved the highest mIoU of 0.8959, while the three Transformer-based models (Ensemble_Transformers, SegFormer, and Swin transformer) achieved the second, third, and fourth highest mIoU scores of 0.8777, 0.8594, and 0.8412, respectively. Testing on seen datasets further demonstrated the superiority of Vision Transformers, with Ensemble_Transformers achieving the highest mIoU scores of 0.8795 on ClinicDB and 0.8478 on Kvasir. Among all other models, Transformer-based models achieved the highest performance. Additionally, on unseen datasets, all Transformer-based models displayed superior generalization ability, where Ensemble_Transformers achieved the highest mIoU scores of 0.7047 on ColonDB, 0.8314 on EndoScene, and 0.6581 on ETIS, significantly outperforming CNN-based models like U-Net (0.5718, 0.7335, 0.4889) and DeepLabV3 (0.5077, 0.7335, 0.6581).

These findings demonstrate that Vision Transformer-based models are more robust and generalize better in polyp segmentation tasks in both seen and unseen datasets, making them suitable for clinical applications. By exploring an underexplored area of medical image analysis, this research advances automated diagnostic tools while contributing to the early detection and prevention of colorectal cancer.

Keywords: Artificial Intelligence, Computer Vision, Polyps.

Table of Contents

List of Tables

List of Figures

List of Definitions of Abbreviations

BC5E	Binary Cross Entropy
CNN	Convolutional Neural Network
CRC	Colorectal Cancer
DLA	Deep Layer Aggregation
FLOPS	Floating-point Operations Per Second
GAN	Generative Adversarial Network
HILA	Hierarchical Inter-level Attention
HVT	Hierarchical Vision Transformer
KNN	K-Nearest Neighbors
IoU	Intersection-over-Union
mIoU	Mean Intersection-over-Union
MLP	Multi-Layer Perceptron
QKV	Query, Key, Value
ReLU	Rectified Linear Unit activation function
ROI	Region of Interest
RPN	Region Proposal Network
SGD	Stochastic Gradient Descent
TanH	Hyperbolic tangent activation function
ViT	Vision Transformer

Chapter One: Introduction

This chapter presents a clinical overview and motivation for the study. It identifies the current research gap, the objectives, and the significance of the study.

1.1 Clinical Overview

A colon polyp is an abnormal growth of cells on the inner colon walls, caused by various factors including genetics and unhealthy lifestyle such as smoking, alcohol consumption, physical inactivity, consumption of red meat and processed meat, and obesity [4] [5]. A combination of these risk factors increases the rate of polyps' formation and increases the risk of polyp recurrence after the initial diagnosis [6]. A colorectal polyp can be categorized into two classes: non-neoplastic and neoplastic. Non-neoplastic polyps are non-cancerous and not harmful [7]. On the other hand, neoplastic polyps are sub-categorized into malignant and pre-malignant polyps. Malignant polyps are cancerous, while pre-malignant are not cancerous, but have the risk of becoming cancerous. However, early detection allows for early medical intervention and polyps removal, which contributes to reducing this risk [8]. Polyps' detection in the large intestine is typically achieved by performing a colonoscopy, where a special tube is inserted in the rectum and a tiny video camera is used to capture videos from inside the colon, along with accessory channels that allow insertion of equipment and fluids to cleanse the colonoscope lense and colonic mucosa [9]. The videos are typically reviewed by specialized clinicians to detect possible polyps, for them to be removed directly during the colonoscopy procedure, or another procedure following the polyp detection colonoscopy.

1.2 Clinical Motivation

Polyps are very common as they occur in about 25% of men and women who are 45 years old and above [10]. Colorectal cancer, the second leading cause of cancer death among both men and women, ranks third most prevalent cancer worldwide [2].

To reduce the risk of polyps turning into colon cancer, early polyps' detection and removal is vital, as it allows for early intervention and consequently, reduces colorectal cancer mortalities [11]. Early diagnosis can increase a patient's survival rate from a low 5-year rate to 90% [12].

Detecting polyps is done by analyzing the visual data generated from colonoscopy, where a specialized clinician is typically responsible for this time-consuming, subjective, challenging process, with a miss rate of 22%-28% [13]. It's challenging due to the large variation of polyps in terms of shape, texture, size, and color, as well as having various polyp-like mimics during colonoscopy [14]. Not to mention that colonoscopy is a demanding, time-consuming, high-cost procedure, that is unpleasant for patients. Considering all the mentioned points, robust analysis needs to be performed on output visual data, to fully utilize the colonoscopy procedure.

To achieve that, computer vision is increasingly being presented as an automated medical diagnosis tool, that makes sense of visual data, in this case, polyps' visual data, the same way doctors and endoscopists do, or even better.

1.3 Research Gap

Colorectal cancer (CRC) is the third-most diagnosed cancer worldwide, and recent studies indicate its burden will continue to grow [15]. However, the mortality rate is significantly

reduced by early polyps detection and medical intervention [16], and for this reason, the continued efforts and investigations in the area of polyps detection is a vital topic. Traditionally, doctors would go through a demanding subjective task of visually analyzing colonoscopy videos to detect polyps.

To address these constraints, researchers adopted hand-crafted methods for image segmentation tasks [17], which consist of two main steps, feature extraction and classification. Those methods achieved relatively low segmentation performance, and for this reason, the need for deep learning techniques exploration emerged. CNNs were introduced as the backbone of segmentation methods, and they have been the dominant deep-learning algorithms in image segmentation applications, for polyp segmentation and classification in colonoscopy images [7] [8] [18] [19]. Recently, Vision Transformers, which are deep learning models that consist of multiple self-attention layers to differentially weigh the significance of parts of the input data, are being introduced as the backbone of image segmentation methods [20] [21] [22]. Studies in implementing polyps' segmentation with Vision Transformer-based methods have also been conducted [23] [24].

Vision Transformers have proven to outperform Convolutional Networks when trained on large datasets [25], however, despite existing research efforts, deploying Vision Transformers in polyps image segmentation tasks is still considered an under-explored area.

1.4 Contribution

This thesis's primary contribution to medical image analysis lies in its systematic examination of Vision Transformers as an innovative approach for polyp segmentation, and their comparison with traditional Convolutional Neural Networks (CNNs). This study compares

Vision Transformer-based models, such as SegFormer and UperNet with Swin Transformer, against traditional CNN-based models like DeepLabV3 and U-Net. Utilizing five widely used datasets (Kvasir, CVC-ClinicDB, CVC-ColonDB, EndoScene, and ETIS). This research evaluates model performance on seen and unseen datasets to gain greater insights into generalization capabilities and the clinical applicability of different approaches. Our findings revealed Vision Transformers as superior to CNNs for handling polyp segmentation challenges such as variations in shape, size, and texture complexities.

Experiments with data augmentation and hyperparameter tuning have revealed critical components that enhance segmentation accuracy, further optimizing the performance of implemented models.

In summary, this thesis advances the field by highlighting the effectiveness of Vision Transformers for polyp segmentation, presenting them as a robust alternative to CNNs.

1.5 Research Objectives

The purpose of this thesis is to investigate Vision Transformers as an innovative approach for semantic segmentation of polyps in colonoscopy images and compare their performance against traditional Convolutional Neural Networks (CNNs). This investigation is driven by the clinical need for accurate and efficient detection of polyps to reduce colorectal cancer risks, with advanced deep learning techniques addressing limitations associated with manual detection such as high miss rates, subjectivity, variations in shape, texture size and color variation of polyps.

This thesis evaluates and benchmarks the segmentation performance of CNN- and Vision Transformer-based models against five widely-used medical datasets (Kvasir, CVC-

ClinicDB, CVC-ColonDB, EndoScene, and ETIS). Performance testing included testing these models on seen as well as unseen datasets to provide an assessment of their generalization capabilities as well as insight into whether or not they would suit real-world clinical applications.

By systematically investigating the effectiveness of Vision Transformers in this context, this thesis adds to an underexplored area of medical image analysis, while seeking to advance automated diagnostic tools aimed at early detection and prevention of colorectal cancer.

1.6 Research Questions

This work aims to utilize Vision Transformers to develop an effective polyp detection system. For that purpose, the following key factors were considered to guide this study:

- How effectively can those deep learning models detect polyps?
- What pre-processing techniques to perform?
- How do we optimize the system's performance?
- What's the system's performance with multiple datasets? i.e., how well does the model generalize?

Targeting those questions helps with the process of designing the polyps' detection system, to help specialists get the most reliable results when put in clinical use. Moreover, developing and successfully adopting this system will get other researchers to invest in designing similar systems for other diseases.

1.7 Thesis Outline

- Chapter 1: Introduction.

This chapter presents an introduction to the study, a clinical overview, and motivation. It highlights the current research gap in the field and outlines the study objectives and significance.

- Chapter 2: Literature Review.

This chapter presents a clinical background of polyps' detection. Existing medical imagery segmentation methods are surveyed, including feature-based, convolutional neural networks-based, and Vision Transformer-based approaches, highlighting their strengths and weaknesses. Then, related work in Vision Transformer-based models and polyps' detection is reviewed.

- Chapter 3: Method.

This chapter presents the implementation methodology pipeline of the proposed architectures. The datasets utilized in the study are introduced, and then each component of the implementation pipeline is discussed, starting with the image preparation step, followed by reporting the training and tuning conducted experiments. The last section highlights testing the models' generalization performance.

- Chapter 4: Results and Discussion.

This chapter reports and analyzes the results of baselines, and the different experiments with the proposed methods, across the implementation pipeline, by analyzing both quantitative and qualitative results.

- Chapter 5: Conclusion and Future Work.

This chapter concludes the thesis by summarizing the key findings and contributions, discussing any limitations to this work, and suggesting areas for future investigation.

Chapter Two: Literature Review

2.1 Introduction

This chapter presents a clinical background on colon polyp classes, colorectal cancer screening procedures, and colonoscopy procedures, including colonoscopy procedures details. Furthermore, relevant studies are explored, and their applications in image segmentation tasks are evaluated for effectiveness.

Colorectal cancer (CRC) is a considerable global health burden, with projections suggesting an increase in future cases. Early detection via colonoscopy allows for intervention to improve survival rates and can save lives.

Polyp detection traditionally relies on visual inspection by experienced medical professionals during and after colonoscopy. Unfortunately, this process can be time-consuming, subjective, and susceptible to human error. This thesis explores the use of deep learning to augment colonoscopy by automating image segmentation for polyp detection.

A comprehensive literature review on deep learning and computer vision techniques is presented, including a detailed review of studies on the topic of polyps' segmentation with colonoscopy imaging. The scope of the review will cover various computer vision methods, including:

- Feature-based methods: This section will examine traditional approaches that rely on hand-crafted feature extraction and classification algorithms.
- Convolutional Neural Network (CNN)-based methods: Since CNNs are the dominant deep learning algorithms in image segmentation tasks, this section will provide a

comprehensive review of CNN architecture, and their applications in various computer vision tasks, highlighting the applications on the topic of this study, semantic segmentation. The advantages and limitations of CNNs in this context will be discussed.

- Vision Transformer (ViT)-based methods: This section will explore the structure and applications of Vision Transformers in various computer vision applications. Also, studies that have integrated ViTs into their semantic segmentation pipelines will be reviewed, while comparing their performance against traditional deep learning architectures such as CNNs.

By reviewing both current and recently emerging techniques, the aim is to assess CNNs strengths and weaknesses, as well as to explore the potential of Vision Transformers, on the topic of semantic segmentation.

2.2 Clinical Background

Colorectal cancer (CRC) is the third most frequently diagnosed cancer cases worldwide, and recent studies estimate its rate will grow 60% by 2030 to over 2.2 million new cases and 1.1 million cancer deaths [15]. As detailed in Table 2. 1, polyps in the colon can be benign, which are non-neoplastic, malignant, which are cancerous, or pre-malignant, which are not cancerous but with the potential to become cancerous. Therefore, early polyp detection and medical intervention is of vital importance for improving survival rates [16].

Table 2. 1 Characteristics of the Different Types of Polyps. Polyps in the colon can be benign, which are non-neoplastic, malignant, which are cancerous, or pre-malignant, which are not cancerous but with the potential to become cancerous.

Feature	Malignant Polyp	Pre-malignant Polyp	Benign Polyp
Description	Cancerous, uncontrolled cell growth	Abnormal, but not uncontrolled growth, can become cancerous	Not cancerous, normal cell growth
Cell growth	Uncontrolled	Abnormal, but not uncontrolled	Normal
Cancerous	Yes	No	No
Spread	Can spread	Does not spread	Does not spread
Neoplastic (potential for cancer)	Yes	Yes	No

Colonoscopy is an integral component of all CRC screening programs, which requires inserting a thin flexible tube equipped with a light and a camera (colonoscope) into a large bowl to inflate and examine all aspects and abnormalities of the colon. Those videos are analyzed by doctors during or after colonoscopy, and accordingly, diagnostic or therapeutic interventions can be performed by doctors, which include hemostasis, biopsy, and polypectomy i.e. polyp removal using small instruments passed through the colonoscope.

2.3 Related Work in Computer Vision

In this section, the different algorithms and approaches adopted in the field of computer vision will be surveyed, highlighting the significant differences among them.

2.3.1 Hand-Crafted Feature-Based Methods

Before deep learning-based computer vision algorithms, researchers would adopt hand-crafted methods, which include two main steps: feature extraction and classification [17].

The feature extraction step includes identifying and extracting relevant image features, which include texture-based, edge-based, or color-based features from frames. In the classification step, machine learning algorithms like random forest or support vector machines [26] segment images by classifying each pixel based on its extracted features. Those methods yield a relatively low-quality segmentation performance and suffer from poor generalization ability [27], and that's where the need for deep learning emerged.

2.3.2 Convolutional Neural Networks-Based Methods

Computer vision algorithms usually rely on Convolutional Neural Networks (CNNs), which are recognized deep learning models inspired by the visual perception process of living creatures [28].

CNNs consist of convolutional, pooling, and fully connected layers. Convolutional layers are intended to learn feature representations of the inputs, where inputs are convolved with learned kernels. The output of this convolution is passed through a nonlinear activation function like sigmoid, tanh, and ReLU functions. The nonlinear activation functions add nonlinearities to CNNs, to detect nonlinear features in them [29].

After convolutional layers come pooling layers. A pooling layer typically sits between two convolutional maps, and it's intended to achieve shift invariance by decreasing feature map resolution. CNN computational load is also decreased by pooling layers as they decrease connections between convolutional layers. Multiple convolutional layers and pooling layers

are used together to extract high-level feature representations. After that comes fully connected layers which generate global semantic data, eventually culminating in the last output layer in a CNN.

Training CNNs to get the optimum CNN parameters, e.g., the weight vectors and bias terms, can be obtained by minimizing an appropriate loss function. One of the common optimization algorithms used to minimize loss functions is stochastic gradient descent (SGD) [30]. There are other techniques to optimize CNN's performance, including data augmentation, regularization, weight initialization, batch normalization, hyperparameters tuning, and model depth and width tuning. However, performance is not the only constraint to consider when designing a CNN, the time and resources needed for processing are critical, since they impose a performance-resources trade-off for CNNs. Considering those training requirements for CNN-dependent algorithms, and instead of designing the architecture from scratch, it's more efficient to utilize transfer learning, by deploying CNN architectures that have been previously trained on very large datasets, and then fine-tuning them for each application, using the application-specific dataset.

In 2012, [31] proposed AlexNet. In AlexNet, the number of feature extraction stages was increased from five in previous structures to seven. It also uses large-size filters (5×5 and 11×11) in the earlier layers. To increase model robustness, AlexNet utilizes dropout, where it randomly passes over several transformational units throughout the training stage. To enhance generalization, local response normalization and overlapping subsampling were performed. Furthermore, to enhance the rate of convergence and reduce the vanishing gradient problem, the activation function ReLU was utilized. Further performance improvements were possible.

In 2014, [32] proposed the Visual Geometry Group-VGG model. VGG models feature more layers than AlexNet, as there are different variations of them, VGG-11, VGG-16, and VGG-19, based on the number of layers. They use smaller filters (3 × 3 filters) rather than the previously used 5 × 5 and 11×11 filters, which produced the same influence as the large-size filters while reducing computational complication. For network simplicity, VGG inserts 1×1 convolutions in the middle of the convolutional layers. VGG models use pooling (max pooling), however, they are computationally expensive because of utilizing around 140 million parameters. To attempt to reduce computational resources, ResNet was introduced.

In [33], ResNet was introduced in 2016, where several types of ResNet models were developed based on the number of layers (the smallest with 34 layers and the largest with 1,202 layers). However, the depth of the model should be thoroughly selected, since the accuracy of the model prediction tends to decrease as the depth increases beyond a certain limit. The innovative idea of ResNet is a reference-based residual learning framework, to address the problem of training a deeper network, where the parameter-free and data-independent residual connections enable cross-layer connectivity using skip connections between layers, and prevent gradient diminishing by accelerating convergence. In comparison with VGG, ResNet has lower computational complexity, even with larger depth.

EfficientNet was proposed in 2019 [34] to further enhance performance. Instead of randomly changing model width, depth, or resolution with random scaling factors, EfficientNet uses a compound coefficient technique that uniformly scales dimensions with a fixed set of scaling factors. This approach works because while increasing one dimension can improve model performance, it is best to balance out their scales. EfficientNet fine-tunes each dimension independently to reach maximum accuracy while penalizing any that are computationally

heavy or slow inference; its authors used AutoML and scaling method to develop seven models that outperformed previous state-of-the-art CNNs by being more accurate while being smaller in size or speed inference.

In image segmentation tasks, fully connected Networks (FCN), as proposed in [35], which were among the first CNN models used in semantic segmentation tasks, leveraged pre-trained image classification models like VGG-16 for feature extraction and applied classification layers to predict pixel-wise labels. However, those architectures suffered from limitations that the DeepLab series attempted to overcome.

DeepLabV1 was proposed in [36], and it was built upon the concept of Fully Convolutional Networks (FCNs) for semantic segmentation. DeepLabV1 offered solutions to two limitations in previous semantic segmentation models. The first solution was to preserve spatial information, as early models would down-sample images leading to loss of spatial details. DeepLabV1 addresses this with atrous convolution - an innovative technique that incorporates an expanded kernel within the convolution kernel to capture wider context around each pixel while maintaining feature map resolution, unlike standard convolution which uses neighboring pixels for calculations, atrous convolution incorporates "holes" within its kernel to analyze wider areas with equal kernel size analysis.

DeepLabV1 offered their solution to refine object boundaries, since earlier models focused primarily on local feature extraction, they introduced a fully-connected conditional random field (CRF) as a post-processing step. The CRF is a probabilistic graphical model that leverages relationships among neighboring pixels in order to refine the segmentation mask, particularly beneficial in improving object boundary accuracy.

DeepLabV2 was proposed in [37]. DeepLabV2 expanded upon the foundations established in DeepLabV1, mainly atrous convolution and fully connected CRFs, but with one key difference: Atrous Spatial Pyramid Pooling (ASPP). ASPP was inspired by Spatial Pyramid Pooling but incorporates atrous convolution for better feature capture across scales within a fixed-size feature map. ASPP utilizes atrous convolutions with various atrous convolution rates on an input image to capture features at various scales; their outputs are then combined into one fixed-size feature map to understand objects of various sizes and their spatial relationships, making the network scale-invariant while processing images of various sizes.

Additionally, DeepLabV2 employs a multi-scale structure that processes images at three scales simultaneously (1.0x, 0.75x, and 0.5x) to achieve multi-scale feature fusion. While this approach increases performance, it also increases computational needs. Finally, DeepLabV2 implements a modified ResNet backbone network to address feature coarsening while expanding the field of view.

DeepLabV3 [38] built on the success of DeepLabV1 and V2, by expanding the ASPP module and mitigating its shortcomings in capturing both global and fine details with atrous convolution. Therefore, DeepLabV3 introduced separate global image pooling within ASPP for a wider context. Additionally, 1x1 convolution on ASPP input preserves fine details, and batch Normalization and dropout regularization improve performance further.

DeepLabV3 goes beyond ASPP by including refinements that enable removing the fully connected CRF post-processing step without impacting performance. A "multi-grid" hyperparameter allows adjusting atrous rates within network blocks for a more comprehensive capture of feature scales. Multiscale training from DeepLabV2 is now

performed during inference time - image processing at various scales then averaged outputs are averaged out for final predictions, although computationally expensive this has improved performance.

The last of the DeepLab series, i.e., DeepLabV3+ [39] contributed with its use of the new backbone, Xception [40], which demonstrated superior results while using fewer computing resources compared with the ResNet backbone used in DeepLabV3. DeepLabV3+ also includes a modified version of Xception with atrous separable convolution, similar to what had been employed with the ResNet backbone used previously.

DeepLabV3+ uses an encoder-decoder architecture, a widely used design in semantic segmentation methods. The encoder, i.e., the Xception backbone, performs image encoding into low-dimensional feature vectors, while the transposed convolution layers in the decoder upsamples the feature vectors through combined convolutions with upsampling to sequentially increase resolution. DeepLabV1 and V2 utilized atrous convolution only as an upsampling technique, while DeepLabV3+ integrates decoder technology to further improve performance.

2.3.3 Vision Transformers Based Methods

Vision Transformers are machine learning models which were proposed in 2017 [41]. They consist of multiple self-attention layers that differentially weigh the significance of parts of the input data, and are lately being introduced in the field of computer vision. They are either used in conjunction with convolutional networks (CNN), e.g., by augmenting feature maps for image classification or by processing the output of CNN networks using self-attention, or by using them to substitute certain aspects of convolutional networks [42], [43]. However,

they don’t need to stay dependent on CNNs, as a transformer applied directly to an image in patches can give exceptionally good performance when used in image classification tasks.

If huge datasets are utilized (14M-300M images) to pre-train the ViT model, and then the model is fine-tuned for a specific task, it can outperform CNNs, as with growing datasets and models, there is no saturating performance [25]. If that’s not applicable, it’s better to use CNNs, since some of the inductive biases inherent to CNNs, like locality and translation equivariance, are missing in transformers [44], and as a result, they won’t generalize well and will tend to overfit. CNNs use pixel arrays, where visual transformers split images into visual tokens (fixed-size patches), embed those flattened tokens including positional embedding as an input to the transformer encoder, then ViT is pre-trained using the large sets of labeled images (supervised training). As with CNNs, after pre-training comes the fine-tuning phase on the dataset of the task under study. However, compared to Vision Transformers, CNNs are easier to optimize, as the performance of a Vision Transformer model depends on many decisions including optimizer, network depth, and dataset-specific hyperparameters.

There has been an increasing number of computer vision tasks utilizing Vision Transformers. One is [25], published in 2021. They directly applied a Transformer to images, with minor modifications. They split the image into patches of fixed size which were linearly embedded, then position embeddings were added, and then the output vectors were fed into a Transformer encoder. For classification, the authors added to the sequence an extra “classification token”. Multilayer Perceptron (MLP) serves as the classification head. They also designed various versions of the ViT by changing the number of layers, hidden size, MLP size, and number of heads, which are ViT-Base, ViT-Large, and ViT-Huge. Using

training datasets of different sizes, they compared the performance of the variations of ViT, CNNs (ResNet and EfficientNet), and hybrid models where they feed the intermediate feature maps into ViT with a patch size of one. To boost performance with small training datasets (ImageNet), they optimize three basic regularization parameters – weight decay, dropout, and label smoothing. However, when trained on large datasets (like JFT-300M), they perform no additional regularization on the smaller subsets and use the same hyperparameters for all settings. This way, the intrinsic model properties are assessed, and not the effect of regularization.

CNN models outperform ViT models when pre-trained on smaller datasets. As data size increases, ViT models outperform CNNs. Larger ViT variants outperform smaller ones as the dataset grows. Hybrids outperform ViT for smaller model sizes but are quite similar as model size increases.

In 2021, [20] proposed the Swin Transformer, a hierarchical Transformer that depends on shifted windowing. Swin-T works by first splitting the input image into patches that do not overlap, using transformer blocks to perform modified self-attention calculations on each patch as individual tokens whose features consist of concatenations of pixel values. As the network becomes deeper, merging layers are used for producing hierarchies while Swin Transformer blocks perform transform features for producing hierarchies with resolutions similar to CNNs. This process of merging and feature-transformation steps repeats several times until an acceptable solution has been reached.

The basic Transformer architecture performs global self-attention. All relationships among patches are calculated, leading to quadratic complexity with the number of patches. Window-

based self-attention does not allow for direct connections between windows, which limits its efficacy as a model. For this reason, to address these challenges and introduce cross-window connections without impacting non-overlapping window efficiency, a partitioning method with shifted windows has been proposed as an alternate arrangement within Swin Transformer Blocks. This approach reduces computational complexity and increases efficiency by restricting self-attention calculations of non-overlapping windows. Its architectural model can be replicated at various scales and its computational complexity is linear with image size.

Authors designed different variations of ViT (ViT-T, ViT-S, ViT-B, ViT L), changing the channel number of the hidden layers, and layer numbers. To get performance results, two training scenarios were tested. When pre-trained on ImageNet-1K, it outperforms previous Transformer architectures, and outperformed previous CNNs, in terms of speed-accuracy trade-off. When pre-trained on ImageNet-22K and fine-tuned on ImageNet-1K, it outperformed previous CNNs, in terms of speed-accuracy trade-off.

In 2021, [22] proposed the Self-distillation with no labels framework (DINO), trying to overcome some of ViTs' drawbacks, which include having high computational demands, needing more training data, and the fact that their features do not exhibit unique properties.

DINO creates a teacher network and a student network, where both networks have the same model architecture, which could be a ViT or a CNN (authors used momentum encoders). In the forward training phase of the model, different crops of the image are created. Then, crop 1 is fed through the student network, while crop 2 is fed through the teacher network. Afterward, a centering and sharpening operation is applied to the output of the teacher

network to avoid collapse. Then, outputs from both networks are normalized using a SoftMax function. To update student network parameters through backpropagation, SGD is used to minimize cross-entropy loss, while DINO uses an exponential moving average on the student network weights to update the teacher network. Self-supervised ViT features explicitly contain the scene layout and object boundaries. They also performed particularly well with a basic k-NN without any fine-tuning, linear classifier, or data augmentation. Also, using smaller patches improves performance.

Compared to CNN-based self-supervised systems, training DINO with ViT significantly reduced computational requirements, while also outperforming CNN-based systems.

2.4 Related Work in Vision Transformers for Semantic Segmentation

In [45], the authors separately use Swin Transformer (SWIN-L) and VOLO (VOLO-D5) as the backbone for extracting video frame features, where both architectures use OCRNet decoder, and the result is the aggregation of the output of the two transformer models through a simple weighted summation on the soft classification result of both models.

For training, data augmentation is done via multi-scale, random crop, random horizontal flip, and metric distortion on images, while they only performed multi-scale and flipping for testing.

After augmentation is performed at the training and aggregation phase, considering Swin Transformer extracts the feature of an image through 4 stages, and to make full use of the output of the transformer without significantly increasing computation, they discard the output of stage 1 and only take the outputs of the last 3 stages of Swin Transformer, resize them to the resolution of the output of stage 2, and perform a concatenating operation. For

VOLO, they used a U-net to combine the last 3 output feature maps, as for VOLO, this architecture gave better results than the resize-concatenate operation.

The performance of the proposed structure gave results better than previous architectures, due to data augmentation (training and testing time), and the aggregation phase as well.

ViT models that were first proposed had fixed-resolution feature maps that proved inefficient when performing dense prediction tasks, such as semantic segmentation of large images. Hierarchical vision transform architecture attempted to address this challenge by combining patch merging, spatial reduction attention, and local shifting windows for efficient self-attention at higher resolutions. Unfortunately, HVT can produce sequential features. Higher-level features are typically created from lower-level features without an inter-level link beyond patch-merging; this poses a problem as localization tends to dominate lower feature maps while semantics is usually found at higher levels. In the past, work has been undertaken to propagate data across features at various levels. Examples of such work include DLA, DenseNet, and D3Net architectures where features propagate only in one direction while higher-level features address potentially noisy or ambiguous lower-level features. HRNet/HRViT encoders introduced convolution connections that work both ways through one fusion module.

Unlike the latest work utilizing fusion models, inter-level connections in [46] use attention and occur iteratively multiple times, allowing a hierarchy to emerge and be refined gradually instead of directly outputted in one step. Hierarchical Inter-Level Attention (HILA) is an attention-based method (standard dot-product attention) that captures bottom-up and top-down updates between features of different levels, as it extends hierarchical Vision

Transformer architectures by adding local connections between features of higher and lower levels to the backbone encoder. In each iteration, higher-level features update lower-level features belonging to them, where semantically similar higher-level features will align the lower-level features to it, and, as a result, the lower-level feature can use this higher-level information to improve and correct its representation. These improved lower-level features are used to re-update the higher-level features, as lower-level features that are semantically aligned to the higher-level features will have more influence when propagated upwards and thus have larger attention weights, which resolves ambiguities in the lower-level features and improves localization for the higher-level features. For the bottom-up updates, HILA propagates features via the self-attention layer from the base HVT model, while in the top-down update, they reuse the attention block in the base transformer architecture to initialize a new attention block. This method can be integrated into the majority of existing HVT architectures [34][24].

Regarding the pretraining phase, they used two versions of the models due to the lack of computational resources to train the larger models, one version has HILA pretrained and one without pretraining, highlighting that pretraining improves performance. To pre-train HILA, the backbone model with HILA is pre-trained from scratch on Imagenet1K, while when skipping pre-training, they reuse pre-training weights provided from the official repository of the backbone model in use, and HILA would be initialized randomly.

Compared to previous state-of-the-art models, even without pretraining, HILA achieved better performance, and pre-trained HILA gives higher improvements. HILA increases performance in terms of both mIoU and F-score, while being significantly more efficient, i.e., fewer parameters and FLOPS.

The semantic segmentation framework in [21] unifies Transformers with lightweight MLP decoders. In their framework, they used a novel positional-encoding-free and hierarchical Transformer encoder. By being positional-encoding-free, the encoder avoids interpolating positional codes when performing inference on images with resolutions different from the training one, which would impact performance. Moreover, the hierarchical part enables the encoder to generate high-resolution fine features and low-resolution coarse features. Also, to improve efficiency, since the main computation bottleneck of the encoders is the self-attention layer, they used a sequence reduction process to reduce the complexity of the self-attention mechanism.

As for the decoder, they used a lightweight All-MLP decoder design that yields a powerful representation without complex and computationally demanding modules. They could use this simple decoder because their encoder has an effective receptive field larger than traditional CNN encoders.

The model first divides the input image into patches of size 4×4, which are inputs to the hierarchical Transformer encoder to obtain multi-level features at different resolutions. After that, it passes these multi-level features to the MLP decoder to predict the segmentation mask.

There is a trade-off between model performance and efficiency, where in this case they are controlled by the sizes of both the encoder and the decoder. As for the encoders, the lightweight model, SegFormer-B0, is efficient while maintaining a competitive performance, which makes it very convenient for real-time applications. However, the largest model, SegFormer-B5, achieves state-of-the-art results on all datasets. Besides encoder size, increasing the channel dimension of the decoder increases performance while decreasing

efficiency, and since the performance plateaus for channel dimensions wider than 768, the channel dimension size of the lightweight decoder is 256 which is used in the two lightweight encoders (SegFormer-B0 and B1), while channel dimension is set to 768 for all other models.

During the training phase, they pre-train the encoder on the Imagenet-1K dataset, and they used various data augmentation techniques with the different training datasets. For simplicity, they did not adopt widely used tricks such as OHEM, auxiliary losses, or class balance loss. Compared to state-of-the-art methods, SegFormer-B0 had better mIoU, outperforming all other real-time methods in terms of parameters, flops, and latency. Moreover, SegFormer-B5 outperforms all other approaches, including SETR, and establishes a new state-of-the-art.

2.5 Related Work in Polyps' Detection

In this section, some of the major contributions in the field of polyp detection will be reviewed.

In 2018, [18] was the first to propose region-based object detection for the polyp detection application. The Region Proposal Network (RPN) uses the feature map from the final convolutional layer of deep-CNN Inception Resnet as input and slides its window onto it. This sliding window is mapped to a feature vector of fixed size, followed by a feature vector mapping layer to a feature vector with fixed dimensions, followed by two layers: (1) box regression to predict locations for proposals (x, y, width, and height); and (2) Box-Classification layer that classifies object scores such as polyps and backgrounds. These output frames and their associated object scores are then sent to detectors along with the original images for processing by deep-CNN, while region proposals undergo "crop-and-

resize" operation to create fixed-size feature vectors. Each vector is then fed through a SoftMax function to estimate class scores and through a regression function to refine proposal coordinates. For training of RPN, positive and negative training samples should be obtained from anchor boxes by computing Intersection-over-Union using ground truth location information. When comparing detection performance between various IoU thresholds, 0.3 or 0.6 should be the negative/positive threshold respectively.

To improve performance and overcome limitations caused by having a small dataset, they applied data augmentation like rotating, scaling, shearing, blurring, and brightening. They also proposed two post-learning methods, automatic false positive learning, and offline learning. To reduce false positives, they collected strong polyp-like FPs which have high detector class scores, then re-trained the detector with the strong FPs, and their proposal coordinates. Deep CNNs are used to automatically learn feature representations through region-based CNNs, unlike conventional object detection techniques such as color wavelet analysis, LBP, or edge detection.

Due to the use of a very deep CNN, their model's frame processing time is 0.39 seconds, which is a disadvantage that needs to be tackled by other studies.

In 2018, [19] proposed pixel-wise localization and frame-wise detection methods, which include both handcrafted and deep learning-based approaches. They addressed two main challenges faced by automatic polyps' detection models: limited datasets and generalizability of models.

They employed GANs [47] for pixel-wise segmentation, which are machine learning algorithms used for unsupervised training that employ two neural networks in a zero-sum game configuration.

They used black border removal, still-frame masking, and a green navigation localizer map for data preparation. As for the small dataset, they handled it using data augmentation techniques, specifically rotation and flipping. Later, they utilized hand-crafted global features (GF) for frame-wise detection, to extract hand-crafted global features such as texture and color distribution, where those features were then combined into a feature vector, before applying a Logistic Model Tree classifier. Other than GF, they experimented with existing deep learning architectures retrained and fine-tuned for frame-wise detection, where all layers from pre-trained CNN were frozen and only the base layers were trained, and then specific layers are unfrozen, and the network is fine-tuned referring to a Bayesian optimization algorithm. Finally, the third approach they assessed for frame-wise detection was GAN, which used a threshold function that took as its input the number of pixels marked positively within a frame; GAN's optimal threshold value was 50 pixels. For block-wise segmentation, they applied RT to the set of input blocks generated via a sliding window and then applied a simple threshold activation function (2 blocks were optimal).

GAN based method outperforms both handcrafted features and approaches based on deep learning architectures with the small dataset used. They achieved a detection accuracy of 90.9% with only 6,356 samples captured by different equipment.

In 2020, [48] fine-tuned a CNN (Darknet19) following YOLOv2 [49] architecture. They applied a median filter with the optimal window size as a post-processing method to reduce

flickering marks (salt and pepper noise), which provided a fair trade-off between recall rate and sensitivity.

To evaluate the generalizability of the algorithm, they validated the model using three independent datasets: 1,338 images with 1,349 polyps, CVC-ClinicDB (612 polyp images), and 7 colonoscopy videos with 26 polyps. They then tested the algorithm's performance using 15 unaltered colonoscopy videos. The algorithm's accuracy was 93.4% for the validation process, and the algorithm detected all the polyps found by endoscopists, along with additional polyps that were not found by endoscopists.

Apart from classification tasks, many visual tasks need the output to include localization, i.e., a class label is supposed to be assigned to each pixel. U-Net architecture was proposed by [50] to achieve the best results.

Two paths are available within this network. One, a contracting path, captures context by applying multiple convolutions followed by rectified line units (ReLUs) and max pooling for downsampling. On the other hand, there exists an expansive path that enables feature extraction, and it performs feature map up-sampling followed by a convolution which reduces feature channels to half, a concatenation with the contracting path' cropped feature map, and convolutions followed by ReLUs. For data augmentation, they used random displacement vectors on a 3x3 grid, with no further pre- or post-processing. U-Net outperformed the sliding-window convolutional network.

In 2022, [23] proposed the SwinE-Net architecture. U-Net CNNs that take polyps as input can struggle to segment them accurately and reliably due to differences in texture, color, size, and shape of polyps. Furthermore, distinguishing between colorectal polyps and background

features may prove challenging. Swin Transformer, on the other hand, requires large datasets and more model training time, with emphasis placed on global context rather than patch-level spatial features, and therefore, SwinE-Net provides a combination of patch-level spatial features with global-level features as a solution.

This architecture refines multilevel feature maps from Swin Transformer and EfficientNet and aggregates them using multi-feature aggregation blocks to produce side outputs that correspond with intermediate polyp masks for both Swin Transformer and EfficientNet (which merge patch-level spatial features with global-level semantic features). Finally, the attentive deconvolutional network-based decoder (Inception-ResNet-based) up-samples feature maps through attention and inception to generate the final polyp segmentation.

The proposed architecture outperformed state-of-the-art previous models and architectures; however, accuracy can be improved further using different sizes of kernels, and different dilated convolutions sizes.

Convolutional neural networks (CNNs), and due to having image noise, camouflage, and polycentric data that is unseen, find it challenging to effectively detect polyp boundaries between polyps. CNNs struggle with evaluating contributions from different levels as well as designing effective mechanisms for merging features from these levels together. To overcome this difficulty,[24] developed an architecture consisting of three models - cascaded-fusion module (CFM), camouflage identification module (CIM), and similarity-aggregation module.

CFM and CIM are used to obtain semantic and locational information on polyps. CIM captures it from low-level features, while CFM captures information from high-level

features. As for SAM, it enables extending these pixel features across an entire polyp area, effectively merging cross-level features.

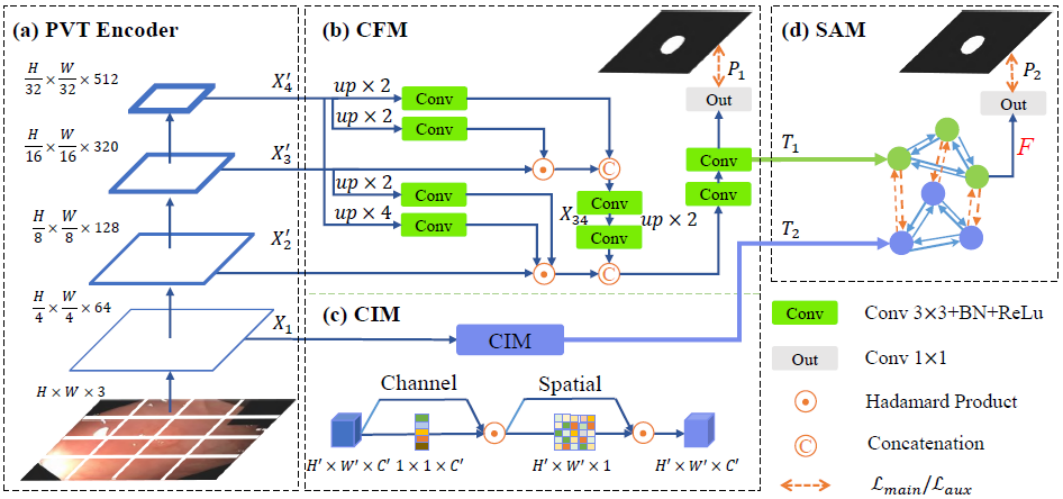

Figure 2. 1 Framework of Polyp-PVT, which consists of a pyramid vision transformer (PVT) (a) as the encoder network, (b) cascaded fusion module (CFM) for fusing the high-level feature, (c) camouflage identification module (CIM) to filter out the low-level information, and (d) similarity aggregation module (SAM) for integrating the high- and low-level features for the final output. This figure is taken from [24].

The authors used the backbone to extract four pyramid features from an input image (x1,x2,x3, and x4) and then adjusted the channel's three high-level features via three convolutional units before feeding these to CFM for fusion to produce T1, while, in parallel, CIM converts low-level features x1 into T2. Later, SAM merges and aligns T1 and T2 together into one feature map F for further processing and prediction of polyp segmentation results in P2 by a convolutional layer. During the training phase, model optimization utilized main and auxiliary loss metrics. Calculation of this main loss between segmentation results

P2 and ground truth allows optimizing polyp segmentation; an auxiliary loss allows monitoring intermediate results generated by CFM.

The authors used the ClinicDB and the Kvasir-SEG datasets to evaluate the learning ability of the proposed model. To evaluate generalization performance, they tested the model on three unseen datasets, ETIS, ColonDB, and EndoScene. Comparing the proposed model with other models, it has a better learning ability, and a good generalization performance, without any pre-/post-processing. One major limitation is the inability to detect accurate polyp boundaries with overlapping light and shadow, causing both FPs and FNs, and that's where image preprocessing could assist the model in the judgment.

As conclusion, Table 2. 2 presents an overview of the previous work that was conducted in polyp detection, where the table outlines the utilized datasets, tasks, algorithms, and best-achieved results.

Table 2. 2 An overview of related work in polyp detection. The table outlines the utilized datasets, tasks, algorithms employed, and best-achieved results.

Citation	Datasets	Task	Algorithm	Results
[18]	Training: CVC-CLINIC ETIS-LARIB Testing: ASU-Mayo Clinic Colonoscopy Video dataset CVC-ClinicVideoDB	Detection (Bounding Box)	R-CNN	precision: 86.5% Recall: 80.8%
[19]	Training: Nerthus Training and Testing: CVC-356 CVC-612 CVC-968 Testing: CVC-12k Kvasir	Segmentation	GAN	detection specificity: 94% accuracy: 90.9%

[48]	Training: Dataset 1 (8,075 polyp images /Private) Dataset 2 (420 polyp images /Private) Validation: Dataset A (1,338 polyp images /Private) Dataset B (CVC-ClinicDB) Dataset C (7 colonoscopy videos /Private) Testing: Dataset D (134 min of colonoscopy videos /Private)	Detection (Bounding Box)	YOLOv2	sensitivity: 96.7% (Dataset A) sensitivity: 90.2% (Dataset B) sensitivity: 87.7% (Dataset C) sensitivity: 89.3% (Dataset D)
[24]	Training and Testing: Kvasir-SEG ClinicDB Testing: ColonDB ETIS EndoScene	Segmentation	Vision Transformer	mIoU: 0.86 (Kvasir) mIoU: 0.89 (ClinicDB) mIoU: 0.73 (ColonDB) mIoU: 0.71 (ETIS) mIoU: 0.83 (EndoScene)
[23]	Training and Testing: Kvasir ClinicDB Testing: ColonDB ETIS EndoScene	Segmentation	CNN and Vision Transformer	mIoU: 0.87 (Kvasir) mIoU: 0.89 (ClinicDB) mIoU: 0.73 (ColonDB) mIoU: 0.69 (ETIS) mIoU: 0.84 (EndoScene)

2.6 Summary

In conclusion, this chapter presented an exhaustive overview of colon polyps, colorectal cancer screening methods, and traditional polyp detection during and after colonoscopy procedures. Early detection through colonoscopy is key for increasing survival rates; however human error and subjectivity may compromise its efficacy. Therefore, this chapter explored deep learning, particularly image segmentation techniques, to serve as an automated polyps detection method.

An in-depth literature review of computer vision techniques was conducted, including image segmentation techniques, while studies on polyps' detection with colonoscopy images were highlighted. This review included various approaches such as feature-based methods,

Convolutional Neural Network (CNN)-based approaches, and Vision Transformers (ViTs), discussing their respective advantages and limitations, particularly the dominant deep learning architecture, CNNs. Likewise, studies that included ViTs were examined for comparison against CNNs.

This review provided a basis for understanding the current landscape of deep learning-based polyp segmentation in colonoscopy images. By assessing existing techniques and their strengths and weaknesses, an opportunity is created for further investigation of ViTs as emerging deep learning models to enhance accuracy, efficiency, and reliability when it comes to polyps' detection.

Chapter Three: Methodology

3.1 Introduction

This chapter presents the methodology followed for medical image semantic segmentation of polyps in colonoscopy images. First, the datasets utilized in this study for models' training and testing are presented and discussed. They are five widely-used datasets (Kvasir [51], CVC-ClinicDB [52], CVC-ColonDB [53], EndoScene [54], and ETIS [55]), which are well-known in computer vision applications within the medical field, as they contain polyp images with their respective ground truth masks for training and evaluating segmentation models. Next, the implementation details and the main implementation components are outlined.

The chapter then elaborates on the deep learning pipeline implemented for this study. This pipeline includes data preparation steps such as data split and data augmenting. Then, the segmentation models are implemented, where DeepLabV3 and U-Net, along with their ensemble model, are implemented as the baseline CNN-based polyp segmentation techniques, and for the contribution of this study, state-of-the-art Vision Transformer models are implemented; SegFormer and UperNet with Swin Transformer, along with their ensemble model, as Vision Transformer-based models for segmenting polyps. RayTune is adopted for hyperparameter tuning to ensure optimal model configuration.

Finally, trained models are evaluated on the testing datasets, both seen and unseen datasets, to assess their generalization ability and suitability for real-world applications or highlight areas for further improvement.

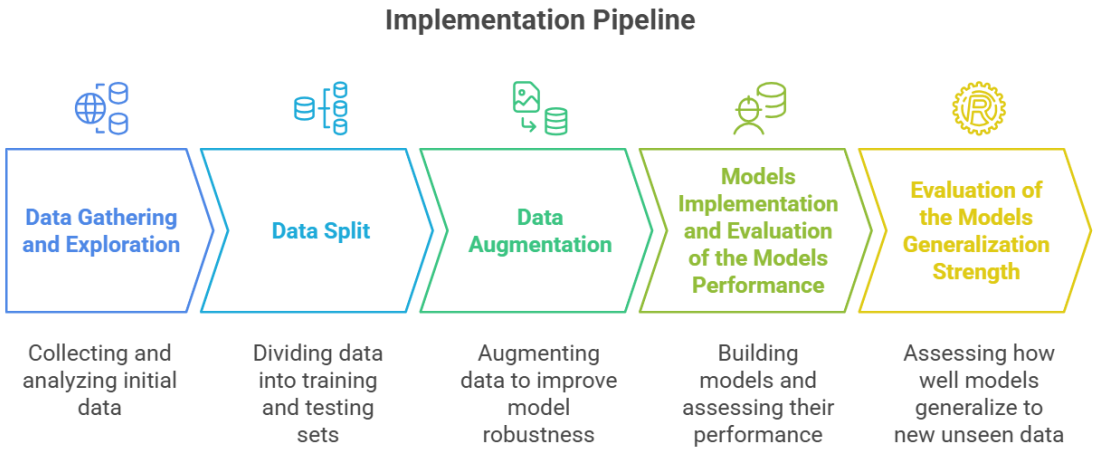

Figure 3. 1 Implementation Pipeline. The main components of the pipeline are data gathering, exploration, split, and online augmentation, model implementation, and performance and generalization evaluation. The models implemented are CNN-based and Transformer-based models. The CNN-based models are U-Net, DeepLabV3, and their ensemble. The Transformer-Based models are SegFormer, Swin Transformer, and their ensemble.

3.2 Datasets Description and Exploration

In this section, the utilized datasets are discussed in detail. In this work, A total of five polyp segmentation datasets are used: Kvasir [51], CVC-ClinicDB [52], CVC-ColonDB [53], EndoScene [54], and ETIS [55], which are well-known in computer vision applications within the medical field. The datasets are images of polyps captured during endoscopy procedures, along with their ground truth, which is a mask verified by specialists to match the location of the polyp in the original image, and the availability of those masks is what makes those datasets very popular, and since one of the main challenges for computer vision models in general, and especially medical imagery, is data availability, mostly of labeled data. In Table 3. 1, we present an overview of the datasets cited above.

Table 3. 1 An Overview of the Datasets Adopted in This Work.

Dataset	Resolution (Pixels)	Samples	Polyp Variation (Size, Shape, Color)	Presence of Artifacts	Annotation Type	Availability
Kvasir-SEG	332×487 to 1,920×1,072	1,000	Wide	Moderate	Pixel-wise Masks	Public
ClinicDB	384×288	612	Moderate	Low	Pixel-wise Masks	Public
ColonDB	500×574	380	Low	Low	Pixel-wise Masks	Public
ETIS	1,225×966	196	Wide	High	Pixel-wise Masks	Public
EndoScene	500×574	60	Moderate	Low	Pixel-wise Masks	Public

In the following sections, we will examine the distinct characteristics of each dataset, emphasizing aspects such as image quantity, resolution, annotation quality, polyp diversity, and visual attributes.

3.2.1 Kvasir-SEG Dataset

The Kvasir-SEG dataset is a part of the larger Kvasir dataset, however, Kvasir-SEG is specifically assembled for polyp segmentation tasks. It contains 1,000 images of high-resolution images with varying dimensions, along with their respective pixel-level ground truth segmentation masks that were annotated and verified by experts. Moreover, the bounding box coordinates of the polyps in the original images are stored in a JSON file. Therefore, the Kvasir-SEG dataset can be used when designing models for the segmentation, detection, localization, and classification of polyps [56] [57].

It contains a wide variety of polyp appearances in terms of size, shape, color, and texture. The images are captured with modern endoscopic equipment and hence are of high quality.

Additionally, they were taken under different lighting and visibility conditions, which enhances robustness. Figure 3. 2 Below is a sample of the Kvasir-SEG dataset.

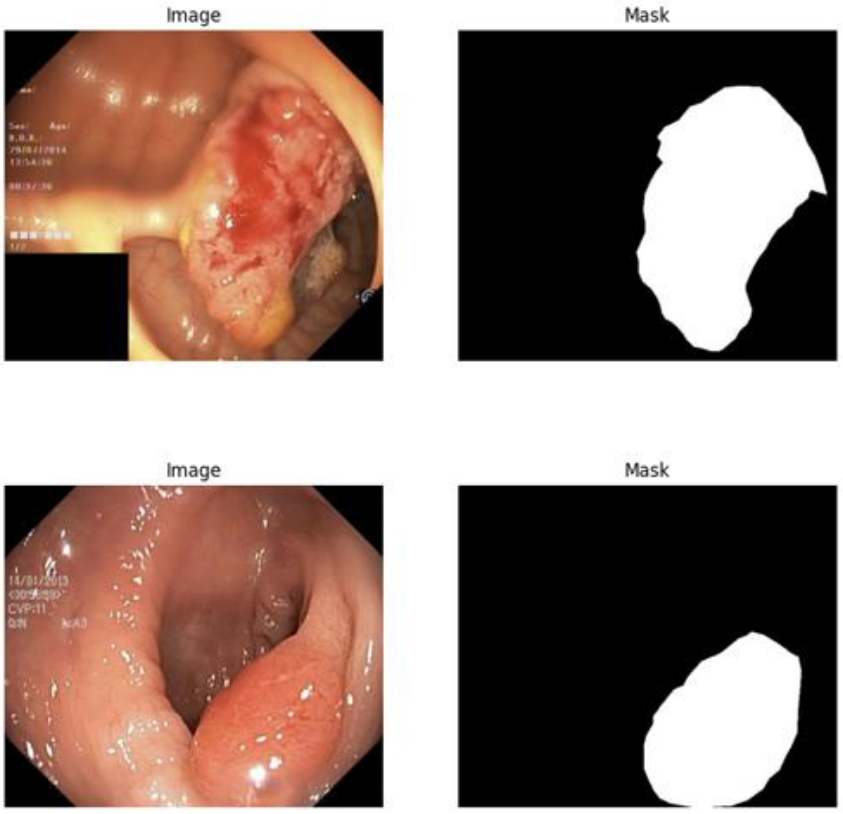

Figure 3. 2 A Sample of the KVasir-SEG Dataset. It includes both the original polyp images and the ground truth masks.

3.2.2 ClinicDB Dataset

Also known as CVC-612, this dataset contains 612 images with a resolution of 384×288 pixels, along with their respective pixel-level ground truth segmentation masks that were annotated and verified by experts. Despite their medium resolution, images are clear with good contrast between polyps and surrounding tissue, and they feature polyps of various sizes and shapes. Additionally, the consistency in imaging conditions reduces complexity during models' training. Figure 3. 3 below is a sample of the ClinicDB dataset.

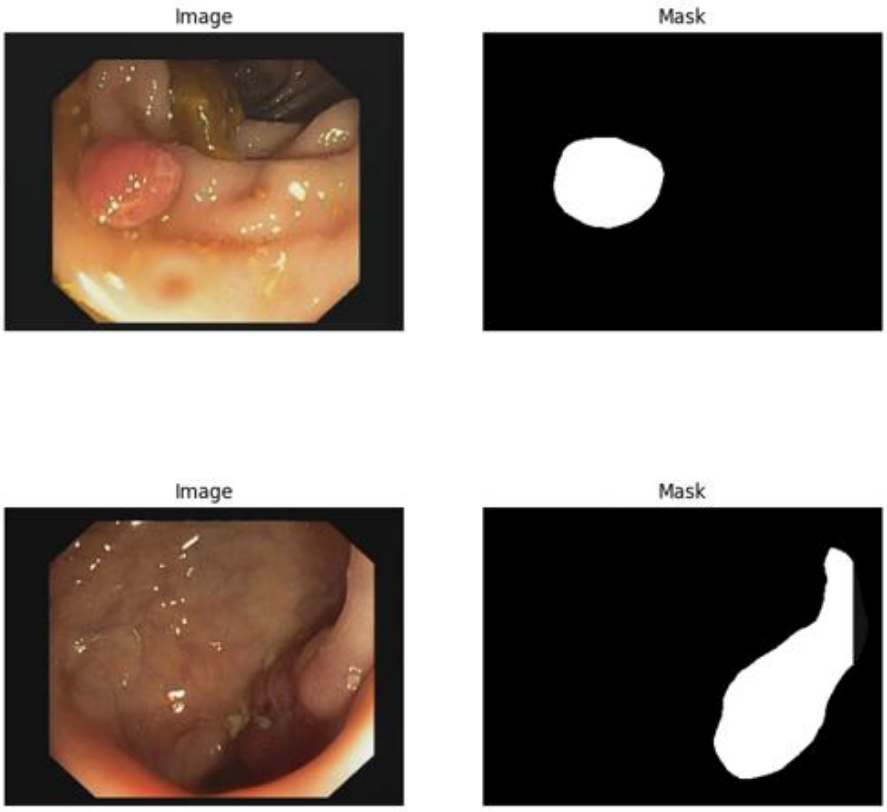

Figure 3. 3 A Sample of CVC-ClinicDB Dataset. It includes both the original polyp images and the ground truth masks.

3.2.3 ColonDB Dataset

ColonDB is a smaller dataset that is aimed at challenging polyp detection scenarios. It contains 380 images, with a resolution of 574×500 pixels, along with their respective pixel-level segmentation masks that were annotated and verified by experts. It contains images with small, flat, or sessile polyps that are harder to detect, while the higher resolution allows for finer detail analysis. As for visual complexity, the backgrounds can be complex with folds and reflections, increasing segmentation difficulty. Overall, the dataset mimics challenging real-world cases encountered in practice. Figure 3. 4 below is a sample of the ColonDB dataset.

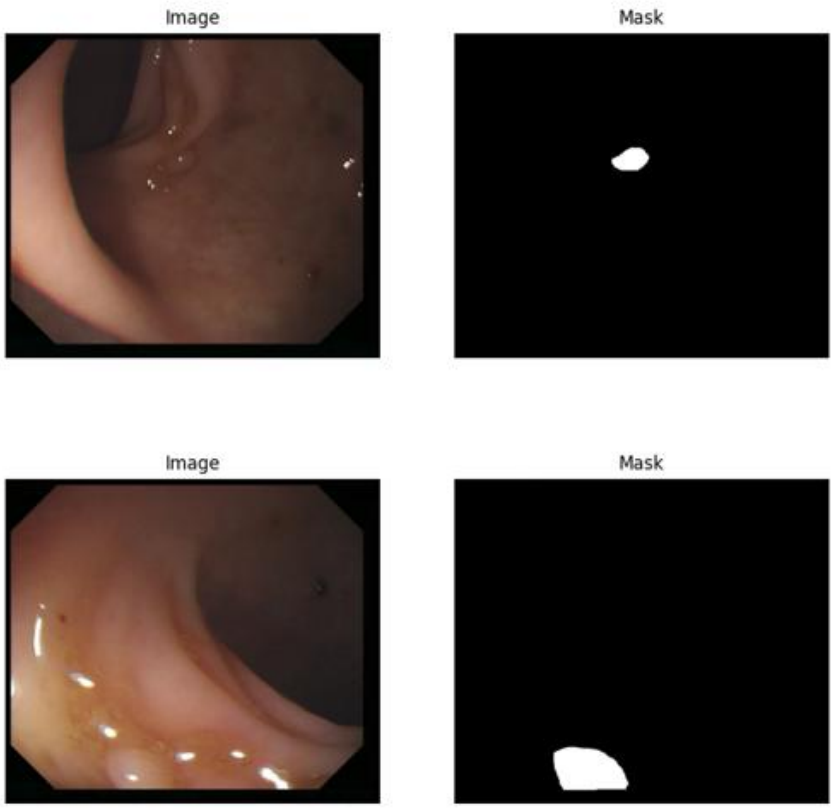

Figure 3. 4 A Sample of CVC-ColonDB Dataset. It includes both the original polyp images and the ground truth masks.

3.2.4 ETIS Dataset

This dataset is designed to test the robustness of polyp detection algorithms under challenging conditions. It includes 196 images with a resolution of 1225×966 pixels, along with their respective high-quality pixel-wise ground truth masks that were annotated and verified by experts. Images include motion blur, noise, low contrast, and poor lighting, and they feature small and subtle polyps, increasing detection difficulty. The high resolution of images requires more computational resources; however, it enables detailed examination and makes it useful for stress-testing algorithms and evaluating performance limits. Figure 3. 5 below is a sample of the ETIS dataset.

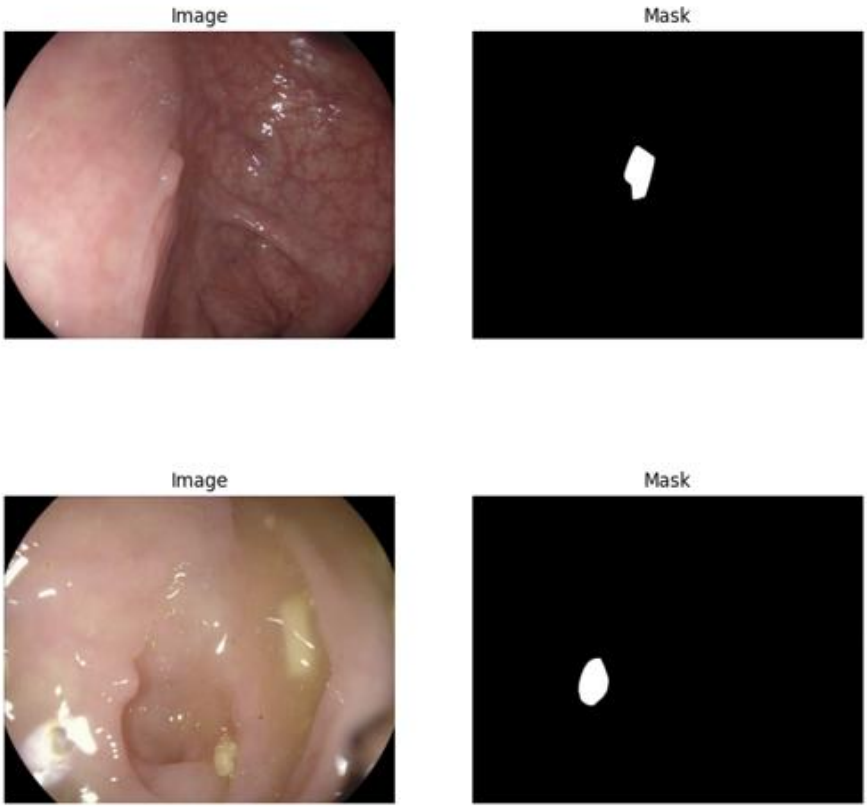

Figure 3. 5 A Sample of ETIS Dataset. It includes both the original polyp images and the ground truth masks.

3.2.5 EndoScene Dataset

The EndoScene dataset focuses on typical polyp appearances in clinical settings. It contains 300 images with a resolution of 500×574 pixels, along with their respective accurate pixel-wise ground truth masks that were annotated and verified by experts. It Includes a moderate range of polyp sizes and types, and balances between image clarity and realistic procedural conditions, which makes it representative of common clinical scenarios. Figure 3. 6 below is a sample of the EndoScene dataset.

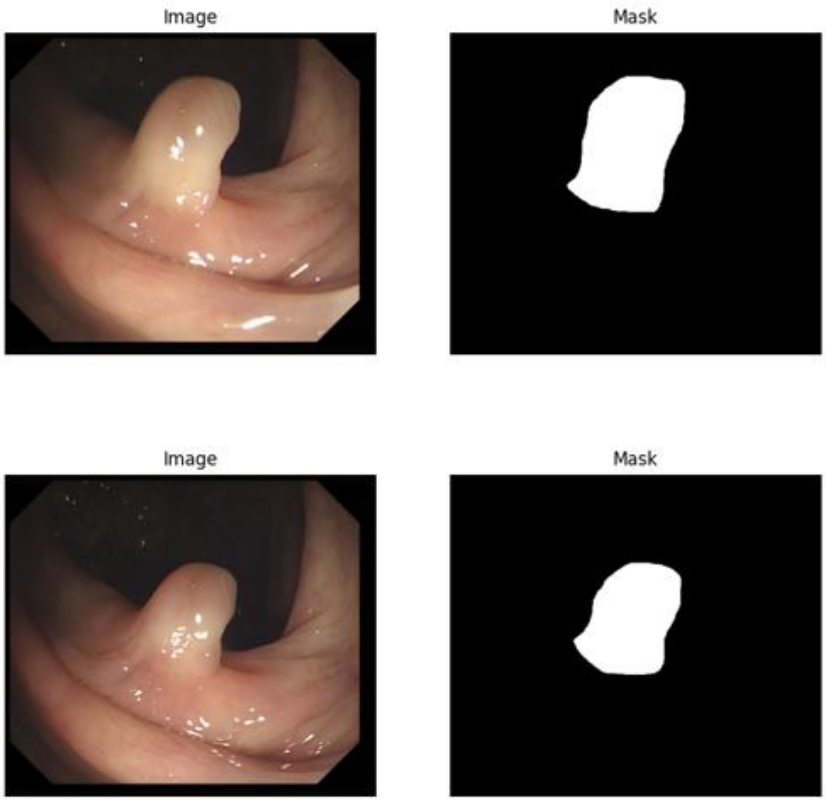

Figure 3. 6 A Sample of EndoScene Dataset. It includes both the original polyp images and the ground truth masks.

3.2.6 Comparative Analysis of the Five Utilized Datasets

To implement a comparative analysis of the datasets, various aspects are considered, which include dataset size, image resolution, image quality, and polyp characteristics variety.

Comparing the size of the datasets, Kvasir-SEG offers the largest dataset, beneficial for training data-intensive models. ETIS has the fewest images, focusing on quality and challenge rather than quantity. As for the resolution and quality of images, ETIS provides the highest resolution, while CVC-ClinicDB has lower resolution images, which may limit fine details but reduce computational load, however, Kvasir-SEG, CVC-ColonDB, and EndoScene have a balance between resolution and practicality. Considering polyp

characteristics, Kvasir-SEG and EndoScene include a broad spectrum of polyp appearances, while CVC-ClinicDB offers a moderate variety, and CVC-ColonDB and ETIS focus on difficult-to-detect polyps, providing a challenge for advanced algorithms. When performing a visual inspection of the datasets, it's observed that Kvasir-SEG images are generally well-lit with high contrast, aiding in visual segmentation, while CVC-ColonDB images may include visual obstructions such as mucus or debris, and ETIS images present visual challenges like poor lighting and noise, simulating less-than-ideal clinical conditions. Finally, all datasets provide high-quality, pixel-level annotations, essential for supervised learning tasks.

In this section, five public medical imagery datasets were explored. All datasets serve as valuable assets for computer vision-based algorithms related to medical imaging; not only due to their size, but also to the accuracy of annotations and masking provided - which subsequently ensure more reliable outcomes of any designed model.

3.3 Implementation Details

The components of the implementation pipeline of our work, including data preparation, data processing, and modeling were all implemented with Python programming language, and executed with the Colab Pro+ platform, due to the high computational needs of our work as a computer vision task. The main components used are the PyTorch framework [58], Albumentations library [59] and RayTune library [60].

- PyTorch is a deep learning framework utilized to implement various deep learning tasks, including computer vision applications that are based on architectures like CNNs and Transformers.

- Albumentations is a powerful Python library that enables performing various image augmentation operations.
- RayTune Python library is mainly utilized for hyperparameter tuning, through state-of-the-art optimization algorithms.

In addition to these core libraries, other Python libraries were utilized, including NumPy [61], Scikit-learn [62], Pandas [63], and Seaborn [64].

- NumPy is a fundamental library for scientific computations and data structuring.
- Pandas library facilitates the efficient handling and manipulation of data.
- Scikit-learn (Sklearn) enables data processing, such as feature scaling and normalization, as well as data splits.
- Seaborn is a data visualization library built on top of the Matplotlib library. It is utilized to create rich and aesthetically appealing data visualizations via a high-level interface.

The above libraries were utilized to execute the end-to-end implementation pipeline designed in this study.

3.4 Experimental Setup

This section presents the deep learning pipeline for medical image segmentation, including data preparation, models' training, hyperparameter tuning, and models' testing. This pipeline was applied to all three CNN-based models (DeepLabV3, U-Net, and their ensemble model), as well as Vision Transformer-based models (Swin transformer, SegFormer, and their ensemble). The pipeline includes data split, online data augmentation, and training with hyperparameter tuning using RayTune. To test the models' performance, they were tested

with the remaining of the training datasets, ClinicDB and Kvasir-SEG. Additionally, to assess how well the models generalize with unseen datasets, they are also tested on the three unseen datasets, ColonDB, ETIS, and EndoScene. The evaluation allows us to measure the efficacy of the deep learning pipeline in segmenting medical images and improving the performance of the segmentation model.

3.4.1 Data Split and Cross-Validation

To be able to compare our findings with other benchmark studies, we adopted the same training strategy used in [23] [24] [65] [66], where the training set includes a total of 1,450 images, with 900 images from Kvasir-SEG, and 550 images from CVC-ClinicDB. The testing set includes the remaining images from the five utilized datasets: the seen testing dataset includes 100 images from Kvasir-SEG, and 62 images from CVC-ClinicDB, while the unseen testing dataset includes 380 images from CVC-ColonDB, 196 images from ETIS-Larib and 60 images from EndoScene.

Table 3. 2 An Overview of the Datasets Train-Test Split Strategy Adopted in This Work.

Dataset	Resolution (Pixels)	Training Samples	Testing Samples
Kvasir	332×487 to 1,920×1,072	900	100
ClinicDB	384×288	550	62
ColonDB	500×574	0	380
ETIS	1,225×966	0	196
EndoScene	500×574	0	60

For validation, following the same methodology adopted in [23] [24], we implemented k-fold validation with k=5, where we split the dataset into 5 folds, and each time we used 4 folds of the dataset as the training set and the remaining one as the validation set.

3.4.2 Data Augmentation

To enhance models' performance, robustness, and generalization, online data augmentation techniques were implemented, to dynamically alter training data through flipping, rotation, and color adjustments and other methods to increase the diversity within the datasets, which exposes the models to wider polyp variations in shape, texture and appearance to ultimately enhance their ability to adapt to real-world scenarios and increase generalization performance.

This was achieved utilizing Torchvision's transforms module, which offered various data augmentation techniques that could be applied to both images and masks before each training epoch. The compose function was used to define the set of transformations that will be applied to each image/mask pair such as horizontal/vertical flips, rotations, affine transformations, and adjustments of brightness/contrast/saturation/hue.

3.4.3 Evaluation Metric

The performance of all segmentation models was evaluated using standard medical image segmentation metrics, i.e. cross-entropy loss [67], dice loss [68], mean Intersection-over-Union (mIoU) [69], and overall accuracy scores.

- Cross-Entropy Loss

Cross-entropy is a widely used loss function used in medical imaging segmentation applications. It measures the difference between the predicted probability distribution of the segmentation mask and the ground truth distribution. When adopting this loss function, a SoftMax activation layer is implemented to convert the raw output logits of the model into a

probability distribution across the segmentation classes. With binary segmentation tasks, the binary cross-entropy loss function can be calculated by the equation below:

$$Loss_{BCE} = -\big(Y\ Log\ \big(Pred(Y)\big) + (1 - Y)\ Log\ \big(1 - Pred(Y)\big)\ \big) \quad (3.1)$$

Where:

Y: The true label (0 for negative class (background), 1 for positive class (object)).

Pred (Y): The predicted probability of the model for the true label (Y).

- Dice Loss

The Dice metric, which is also known as the Dice Similarity Coefficient, is widely used in image segmentation tasks including semantic segmentation. It provides a score based on the similarity between the predicted segmentation, and the actual ground truth, with a value ranging from 0 to 1, where 0 represents no overlap and 1 represents perfect overlap. Dice loss can be represented as:

$$Loss_{Dice} = \frac{2 \times |A \cap B|}{|A| + |B|} \quad (3.2)$$

Where:

A is the ground truth mask

B is the predicted mask

$|A \cap B|$ is the area of the intersection

$|A| + |B|$ represents the total class areas in both the ground truth and the predicted masks

Typically, the Dice loss is used in conjunction with other loss functions like cross-entropy to strike a balance between different aspects of the segmentation task, which was implemented in this work. Where the utilized combined loss can be represented as:

$$Loss_{combined\_BCE\_Dice} = \frac{Loss_{Dice} + Loss_{BCE}}{2} \quad (3.3)$$

- Mean Intersection Over Union

Intersection Over Union is a key metric to evaluate image segmentation model performance. It measures the degree of overlap between the predicted segmentation mask and ground truth annotation. mIoU is the ratio of the area of intersection of the ground truth annotation and the predicted mask, to the area of area on their union. Based on IoU calculation, the Mean Intersection over Union (mIoU) for a set of images is calculated by averaging the IoU values of all images. The IoU for the i-th image can be represented as:

$$IoU_i = \frac{Area\ of\ Intersection_i}{Area\ of\ Union_i}, \quad (3.4)$$

Here, $Area\ of\ Intersection_i$ is the area of the overlap between the ground truth and the predicted masks for image i, and $Area\ of\ Union_i$ is the total combined area covered by both the prediction and ground truth for that image. Finally, the mIoU is calculated by averaging the IoU scores across all N images:

$$mIoU = \left(\frac{1}{N}\right) \Sigma\ (IoU_i), \quad (3.5)$$

A higher mIoU value indicated a better alignment between the model's predictions and the true object mask.

- Overall Accuracy Score

This metric simply measures the percentage of correctly classified pixels within an image. Unfortunately, when dealing with datasets that tend to have dominant background classes, overall accuracy without considering other metrics like mIoU can be misleading. For binary semantic segmentation models, accuracy can be represented as:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}, \tag{3.6}$$

Where:

TP: the number of pixels from the positive class (object) predicted correctly by the model.

FP: the number of pixels from the negative class (background) predicted incorrectly as a positive class (object) by the model

TN: the number of pixels from the negative class (background) predicted correctly by the model.

FN: the number of pixels from the positive class (object) predicted incorrectly as a negative class (background) by the model.

3.4.4 Implementing CNN-Based Models

3.4.4.1 DeepLabV3

To provide an in-depth comparison between transformer-based architectures and CNN-based architectures in image segmentation tasks, the DeepLabV3 algorithm was implemented. The pre-trained model, ResNet-101, was the backbone of the segmentation model due to its strong feature extraction capabilities.

DeepLabV3 is a deep learning architecture optimized for semantic segmentation, employing atrous convolutions to capture multi-scale contextual information while maintaining spatial resolution. To maximize segmentation accuracy, its Atrous Spatial Pyramid Pooling (ASPP) module applies parallel atrous convolutions with variable dilation rates to capture features at multiple scales while also taking into account global context to increase segmentation accuracy.

This architecture serves as an effective basis for comparing with vision transformers for polyp segmentation tasks due to its ability to extract both local and global features for the accurate detection of polyps of all shapes and sizes.

To present a fair comparison between the transformer-based models implemented in this study (SegFormer, Swin Transformer, and their ensemble model), as well as the CNN-based models implemented (DeepLab, U-Net, and their ensemble), our experimental setup was consistent across all models. All training pipeline components remained the same across models: data split and preprocessing, online augmentation techniques, and model training were kept identical to accurately assess each model's strengths and weaknesses in this segmentation task.

$Loss_{combined}$, which is the average of BCE loss and Dice loss, was used to measure discrepancies between predicted and ground truth masks, while for optimizing model parameters, Adam optimizer was utilized. Model training was achieved with the hyperparameters in Table 3. 3, while tracking validation accuracy, mean Intersection over Union (mIoU), and loss values across the training and validation dataset, which includes a total of 1,450 images, with 900 images from Kvasir-SEG, and 550 images from CVC-ClinicDB.

Table 3. 3 DeepLab Training Hyperparameter Tuning Space.

Hyperparameter	Hyperparameter Value
Batch Size	[4,8]
Number of Epochs	[40,70,100]
Learning Rate	[0.0001,0.001]

After the training experiments with the training dataset, and identifying the model with the best results according to k-fold validation results, the performance of this model is evaluated using the testing datasets that include the remaining images from the five utilized datasets: the seen testing dataset includes 100 images from Kvasir-SEG and 62 images from CVC-ClinicDB, and the unseen testing dataset, which is utilized to test the generalization ability of the model, includes 380 images from CVC-ColonDB, 196 images from ETIS-Larib and 60 images from EndoScene.

3.4.4.2 U-Net

The U-Net model is a convolutional neural network architecture. It employs a symmetric encoder-decoder structure. The encoder captures spatial features of an input image using

successive convolutional layers and max-pooling operations to reduce spatial dimensions while increasing depth feature maps. Meanwhile, its counterpart, the decoder, reconstructs segmentation masks by gradually up-sampling feature maps using skip connections, thus maintaining spatial details lost during down-sampling processes, thus providing precise boundary delineation during segmentation tasks.

U-Net makes a suitable baseline model for polyp segmentation models due to its ability to handle limited training data effectively, which is an often difficult challenge in medical imaging. Furthermore, its skip connections ensure fine details such as polyp boundaries are captured precisely, which is something essential in this application. Furthermore, its end-to-end fully convolutional nature ensures efficient computational processing that provides accurate pixel predictions necessary for the precise segmentation of polyps.

U-Net architecture was implemented as another CNN-based architecture, utilized in an image segmentation task. The U-Net architecture we implemented incorporates a pre-trained ResNet152 model as the encoder backbone to enhance feature extraction from input images.

The training pipeline components remained the same as with the other models: data preprocessing and data split, online augmentation techniques, and model training, to accurately assess each model's strengths and weaknesses in this segmentation task.

$Loss_{combined}$, which is the average of BCE loss and dice loss, was used to measure discrepancies between predicted and ground truth masks, and the model training was achieved with the hyperparameters in Table 3. 4, while tracking validation accuracy, mean Intersection over Union (mIoU), and loss values across the training and validation dataset,

which includes a total of 1,450 images, with 900 images from Kvasir-SEG, and 550 images from CVC-ClinicDB.

Table 3. 4 U-Net Training Hyperparameter Tuning Space.

Hyperparameter	Hyperparameter Value
Batch Size	[8,16]
Number of Epochs	[40,70,100]
Learning Rate	(0.0001, 0.01)

After the training experiments with the training dataset, and identifying the model with the best results according to k-fold validation results, the performance of this model is evaluated using the testing datasets that include the remaining images from the five utilized datasets: the seen testing dataset includes 100 images from Kvasir-SEG, and 62 images from CVC-ClinicDB, and the unseen testing dataset, which is utilized to test the generalization ability of the model, includes 380 images from CVC-ColonDB, 196 images from ETIS, and 60 images from EndoScene.

3.4.4.3 Ensemble Model of CNN-Based Models

To leverage the capabilities of the two CNN-based models: DeepLabV3 and U-Net, a semantic segmentation ensemble model was implemented. The ensemble model is constructed by integrating the logits generated by each individual model, weighted by user-specified weights, and combining them. This approach aims to leverage the strengths of both models, potentially providing an enhanced segmentation performance.

$$Logits_{Ensemble\_CNNs} = \frac{\left(W1 * Logits_{DeepLabV3} + W2 * logits_{U-Net}\right)}{(W1 + W2)} \qquad (3.7)$$

$Logits_{Ensemble\_CNNs}$ is the logits generated by the ensemble model, $Logits_{DeepLabV3}$ is generated by the DeepLabV3 model, and $logits_{U-Net}$ is generated by the U-Net model; and $W1$ and $W2$ represent user-specified weights for each model used, where we experimented and evaluated the results with multiple weights, to assign the weights with optimal results.

The performance of the models was evaluated with k-fold cross-validation using the training dataset, and the model with the best results was then tested with the testing datasets: the seen testing dataset that includes 100 images from Kvasir-SEG, and 62 images from CVC-ClinicDB, and the unseen testing dataset, which is utilized to test the generalization ability of the model, that includes 380 images from CVC-ColonDB, 196 images from ETIS-Larib and 60 images from EndoScene.

3.4.5 Implementing Transformer-Based Models

3.4.5.1 SegFormer

SegFormer is a transformer-based model for semantic segmentation that combines global context-capturing abilities with lightweight and efficient design. Its encoder extracts hierarchical multi-scale features using overlapping patch embeddings without needing position embeddings, then its MLP decoder fuses these features to produce high-quality segmentation masks.

SegFormer excels at polyp segmentation due to its ability to capture global and local features that help detect subtle or irregular polyp structures. Furthermore, its efficient architecture guarantees scalability for large datasets while maintaining accuracy, making it a strong candidate when compared with conventional CNN-based models like U-Net and DeepLabV3.

The largest pre-trained version of the SegFormer models, SegFormer-B5, was used to obtain superior performance with manageable complexity. As for the SegFormer model configurations, all segmentation configuration values are kept as the default.

The Hugging Face library was utilized to load the SegFormer model and fine-tune it with our dataset. During training, the model was periodically evaluated to compute the validation loss, mIoU, and mean accuracy across the training and validation dataset, which includes a total of 1,450 images, with 900 images from Kvasir-SEG, and 550 images from CVC-ClinicDB. These values were then reported to RayTune to select the best hyperparameters, by searching the hyperparameter space, which included batch size, number of epochs, and learning rate, as outlined in Table 3. 5.

Table 3. 5 SegFormer Hyperparameter Tuning Space.

Hyperparameter	Hyperparameter Values
Batch Size	[4, 8]
Number of Epochs	[40, 70, 100]
Learning Rate	[0.0001, 0.001]

Finally, after the training experiments with the training dataset, and identifying the model with the best results according to k-fold validation results, the performance of this model is evaluated using the testing datasets that include the remaining images from the five utilized datasets: the seen testing dataset includes 100 images from Kvasir-SEG and 62 images from CVC-ClinicDB, and the unseen testing dataset, which is utilized to test the generalization ability of the model, includes 380 images from CVC-ColonDB, 196 images from ETIS-Larib and 60 images from EndoScene.

3.4.5.2 Swin Transformer

The Swin Transformer is a hierarchical vision transformer that uses a shifted window mechanism to calculate self-attention within non-overlapping windows, thus reducing computational complexity while still capturing long-range dependencies. Additionally, its hierarchical design enables multi-scale feature extraction by combining the strengths of transformers and convolutional architectures; additionally, its shifted windows strengthen connections between regions for improved detail preservation.

This architecture is efficient for polyp segmentation due to its ability to manage irregular shapes and sizes while simultaneously capturing both local and global features. Furthermore, its ability to process high-resolution medical images makes it a robust backbone for segmentation models, offering precise results even during challenging segmentation tasks.

For implementing the Swin Transformer, the smallest pre-trained model, Swin-T, was utilized as the backbone of the segmentation model, and as the decoder, the UperNet decoder was incorporated with the Swin Transformer. The configuration of the backbone model played a crucial role in achieving accurate segmentation results, and the model was implemented with the default configuration settings.

Finally, for hyperparameter tuning, the same methodology that was adopted with SegFormer was followed, by utilizing RayTune to select the best hyperparameters, by searching the hyperparameter space, which included batch size, number of epochs, and learning rate, as outlined in Table 3. 6, and with the same loss function used with other models, $Loss_{combined}$, to be able to compare the performance of the Swin Transformer-based model with the other transformer-based and CNN-based models.

Table 3. 6 Swin Transformer Hyperparameter Tuning Space.

Hyperparameter	Hyperparameter Values
Batch Size	[4, 8]
Number of Epochs	[40, 70, 100]
Learning Rate	[0.0001, 0.001]

After the training experiments with the training dataset, and identifying the model with the best results according to k-fold validation results, the performance of this model is evaluated using the testing datasets that include the remaining images from the five utilized datasets: the seen testing dataset includes 100 images from Kvasir-SEG and 62 images from CVC-ClinicDB, and the unseen testing dataset, which is utilized to test the generalization ability of the model, includes 380 images from CVC-ColonDB, 196 images from ETIS-Larib and 60 images from EndoScene.

3.4.5.3 Ensemble Model of Transformer-Based Models

We developed a semantic segmentation ensemble model by combining the output logits from SegFormer and UperNet with the Swin Transformer backbone. We assigned user-specified weights to each model's logits and combined them to produce a final segmentation prediction. This strategy seeks to benefit from the complementary strengths of the individual models, potentially leading to enhanced segmentation results.

The combination of logits from both models is represented as below:

$$Logits_{Ensemble\_Transformers} = \frac{\left(W1 * Logits_{SegFormer} + W2 * Logits_{Swin}\right)}{(W1 + W2)} \qquad (3.8)$$

$Logits_{Ensemble\_Transformers}$ is the logits generated by the ensemble model, $Logits_{SegFormer}$ is generated by the SegFormer model, and $Logits_{Swin}$ is generated by the Swin model; and

$W1$and $W2$ represent user-specified weights for each model used, where we experimented and evaluated the results with multiple weights, to choose the model with the weights that achieved optimal results.

The performance of the model with the best results was evaluated with k-fold validation using the training dataset, and it was then tested with the testing datasets: the seen testing dataset includes 100 images from Kvasir-SEG and 62 images from CVC-ClinicDB, and the unseen testing dataset, that is utilized to test the generalization ability of the model, includes 380 images from CVC-ColonDB, 196 images from ETIS-Larib and 60 images from EndoScene.

3.4.6 Models' Generalization on Unseen Dataset

In machine learning, models need to be evaluated based on their performance with unseen datasets, i.e. not used during training, which is not only an indicator of their generalization ability but also a method to detect any overfitting in the model. The goal is to evaluate the generalization performance of all models on the unseen dataset which includes 380 images from CVC-ColonDB, 196 images from ETIS-Larib, and 60 images from EndoScene. This evaluation demonstrates the models' abilities to detect hidden patterns in new real-world data and hence indicates whether the model is suitable for deployment in practical applications. On the other hand, poor performance on those test datasets could mean the model is overfitted to the training dataset and requires additional optimization and enhancement to increase its generalization performance.

3.5 Summary

In conclusion, this chapter provided a comprehensive deep-learning pipeline used for medical image segmentation of polyps. The implementation pipeline consists of data preparation

steps like splitting and online augmentation, followed by the training and tuning of the segmentation models. The last component of the pipeline was testing the generalization performance of the models with both seen and unseen datasets. First, the five datasets that were utilized in the models' training and evaluating process, Kvasir, CVC-ClinicDB, CVC-ColonDB, EndoScene, and ETIS, were explored. Later, as a baseline, CNN-based segmentation models were implemented (DeepLabV3 with the ResNet101 backbone, U-Net, and an ensemble of the two). Then, state-of-the-art Vision Transformer models, which are SegFormer, and UperNet with the Swin Transformer backbone, and their ensemble, were utilized. Throughout the training process, RayTune hyperparameter tuning further optimized the models' performance. Finally, the model with the best results according to k-fold validation results was evaluated with the testing datasets that include the remaining images from the five utilized datasets, providing valuable insights on the models' generalization ability for real-world applications and future research directions.

Chapter Four: Results and Discussion

4.1 Introduction

This chapter evaluates the performance of the various deep learning models implemented for our task, wherein the outputs of each component in the study's implementation pipeline are studied, as shown in Figure 3.2, beginning with data preprocessing, data split, and data augmentation techniques, followed by training and tuning of the segmentation models, and lastly testing their performance on the testing datasets.

This chapter details the implementation of four image segmentation models: CNN-based models (DeepLabV3, U-Net, and their ensemble model), and transformer-based models (SegFormer, Swin Transformer, and their ensemble). Each implementation process is described, from data augmentation creation, and hyperparameter tuning strategies used to optimize performance, to the evaluation of the performance of the models based on mean Intersection over Union (mIoU) with the testing dataset that includes both seen and unseen datasets to evaluate segmentation effectiveness and robustness of the models.

This chapter not only presents quantitative results comparing models' performances on the testing polyp segmentation datasets but also offers qualitative analysis for further insight. By studying individual images' behavior of the models, the purpose is to identify each model’s strengths and weaknesses such as generalization ability or presence of false positives - thus offering a thorough evaluation of whether Vision Transformer-based models are suitable for polyp segmentation tasks.

4.2 Data Split and Cross-Validation

From the 5 datasets, the training set includes a total of 1,450 images, with 900 images from Kvasir-SEG, and 550 images from CVC-ClinicDB. The testing set includes a total of 798 images from the remaining images from the five utilized datasets: the seen testing dataset includes 100 images from Kvasir-SEG, and 62 images from CVC-ClinicDB, while the unseen testing dataset includes 380 images from CVC-ColonDB, 196 images from ETIS-Larib and 60 images from EndoScene. Figure 4. 1 and Figure 4. 2 below represent the data split and the resulting training and testing datasets distribution. For training and validation, we implemented k-fold cross-validation with k=5. By splitting the dataset into multiple subsets (folds) and training and validating models using combinations of these folds, k-fold cross-validation ensures that every data point will be used both for training and validation at least once. This process provides a more accurate estimation of model performance by reducing variance caused by single, arbitrarily split data points that do not accurately represent their underlying distribution. Cross-validation also reduces the risk of overfitting to any specific train-test split, producing more generalizable models. When working with limited data sets as in this work, k-fold cross-validation is a particularly efficient technique to use as it maximizes the use of the dataset.

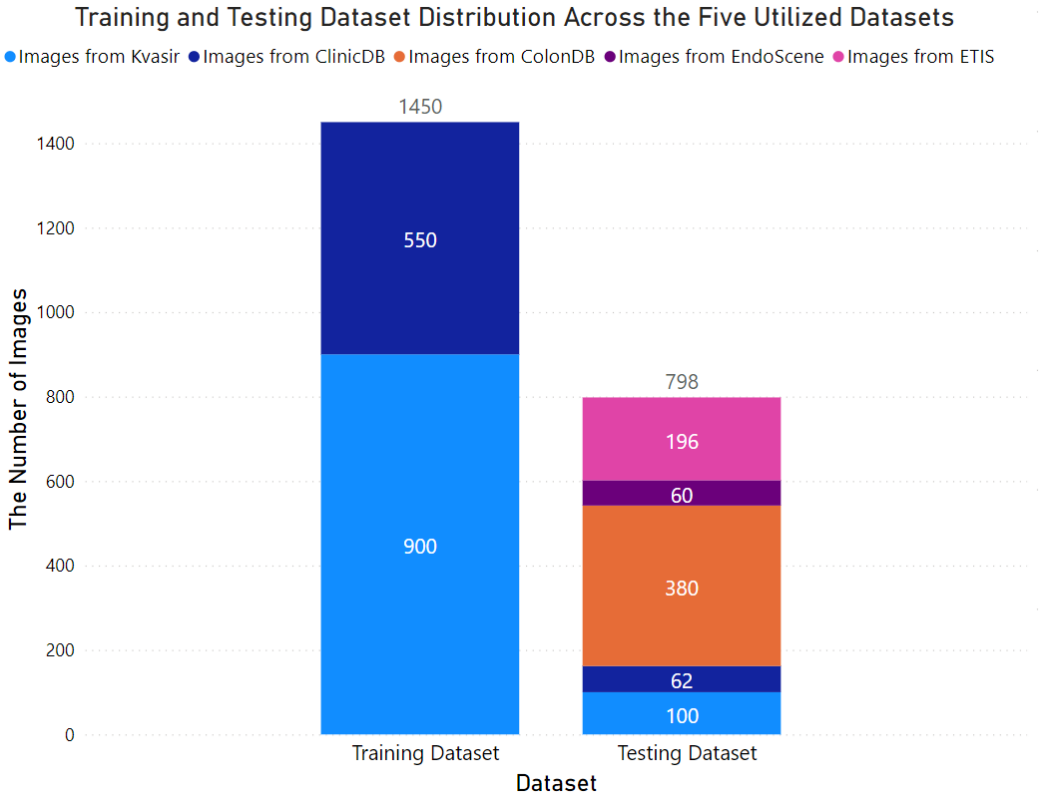

Figure 4. 1 The Details of the Images in the Training and Testing Datasets.

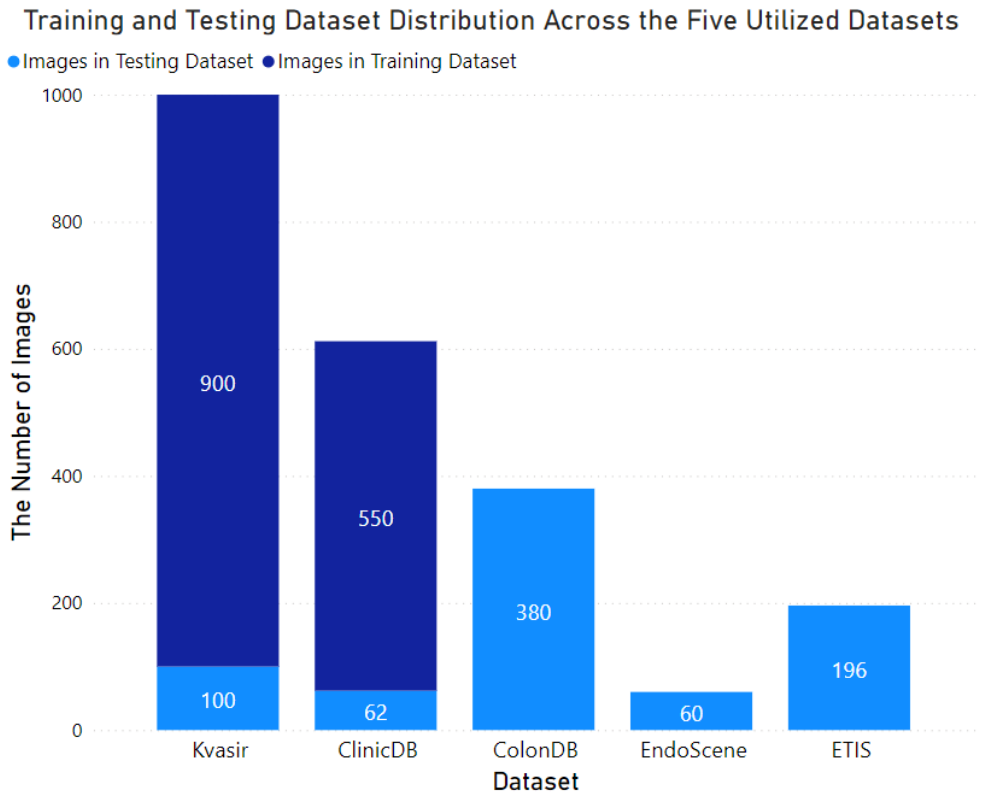

Figure 4. 2 The Split of Each of the Five Datasets, Between the Training and Testing Datasets.

4.3 Data Augmentation

Data augmentation techniques add new variations of the datasets, to expand and diversify them, which enhances deep learning models' performance while avoiding overfitting. Particularly, online augmentation exposes the model to a wide variety of data, when performing augmentation on the fly during training, which further improves model performance and supports avoiding overfitting, while reducing storage requirements. Those techniques are considered crucial for training machine learning models handling medical images to identify structures of interest with optimal performance.

Torchvision's Transforms module was used to perform online data augmentation since it offers various data augmentation techniques that could be applied to the datasets. Using the Compose function, a sequence of transformations is determined for each image and mask to go through. These transformations are:

- Random horizontal flips and vertical flips, to help the model generalize to images with different orientations.

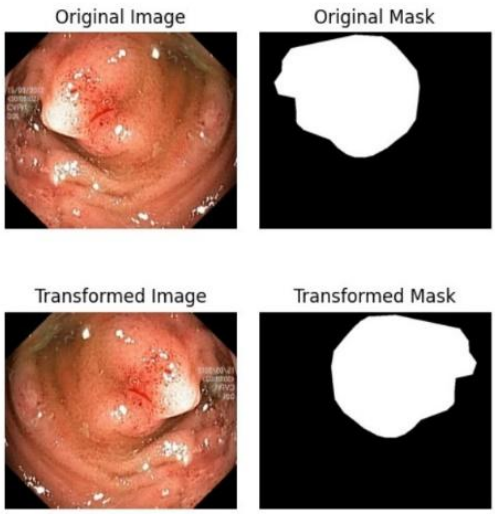

Figure 4. 3 Results of horizontal Flip implemented for data augmentation. It is performed on Images and Masks from the utilized datasets.

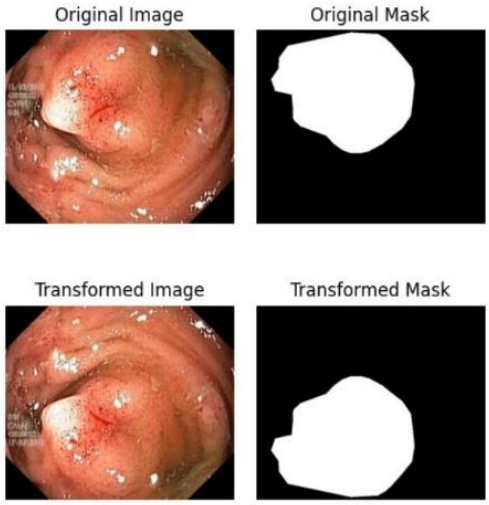

Figure 4. 4 Results of vertical Flip implemented for data augmentation. It is performed on Images and Masks from the utilized datasets.

- Rotations, to help the model generalize to images with different angles of view.

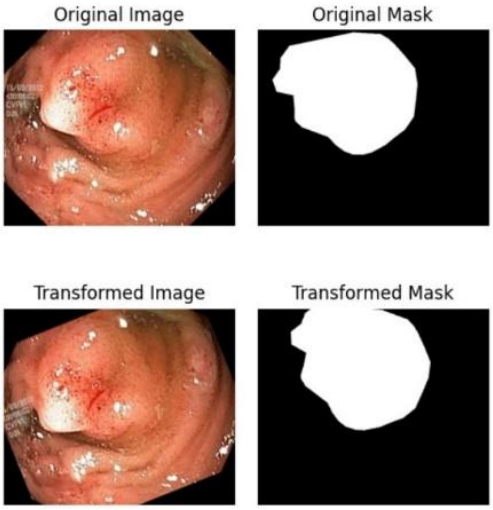

Figure 4. 5 Results of rotation implemented for data augmentation. It is performed on Images and Masks from the utilized datasets.

- Affine transformations, to help the model generalize to images with different scales and aspect ratios.

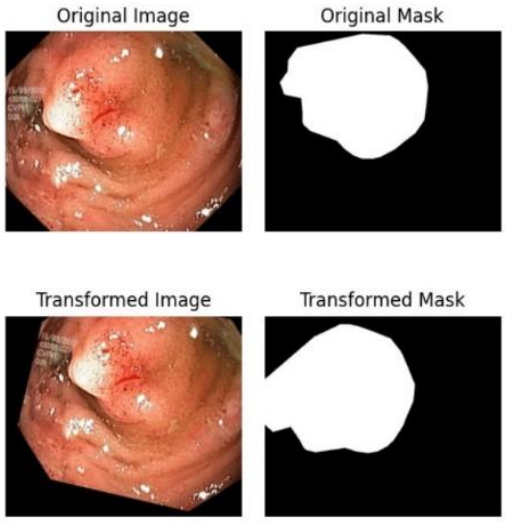

Figure 4. 6 Results of affine Transformation implemented for data augmentation. It is performed on Images and Masks from the utilized datasets

- Adjustments to brightness, contrast, saturation, and hue, which could help the model generalize to images with different lighting conditions and color schemes.

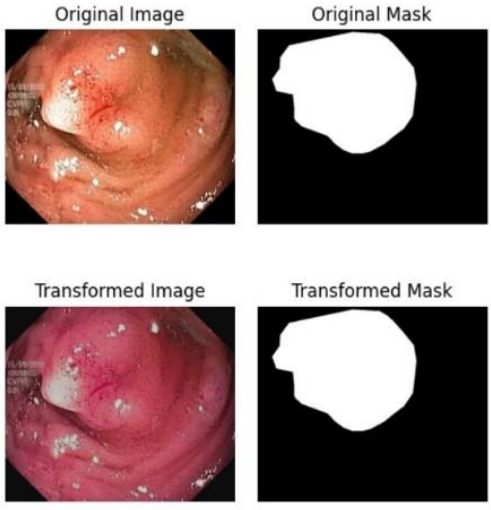

Figure 4. 7 Results of color jitter transformation implemented for data augmentation. It is performed on Images and Masks from the utilized datasets.

4.4 Results of CNN-Based Models

4.4.1 DeepLabV3

As part of the investigation comparing transformer-based image segmentation models with CNN-based image segmentation models, DeepLabV3 with the ResNet-101 backbone was implemented as a CNN-based image segmentation model.

To facilitate an accurate comparison between the implemented transformer-based models, and the CNN-based models, including the DeepLabV3-based model, a consistent experimental setup was maintained, such as using identical data preprocessing techniques, augmentative techniques, and train-validation-test dataset splits for all models.

For training, we implemented k-fold cross-validation with k=5, with the combined loss function ($Loss_{combined\_BCE\_Dice}$), and employed Adam optimizer. Online image augmentation was performed, where a random set of transformations are applied to the input data before each training epoch, to expand the dataset size and variety, to eventually increase the model's performance, robustness, and generalization ability.

RayTune was utilized For hyperparameter tuning, and Table 4. 1 outlines the results of several trials of training the model. The hyperparameters that were varied are batch size, number of epochs, and learning rate, and the evaluation metrics used to compare the different trials are $Loss_{combined\_BCE\_Dice}$ and mIoU.

Among the different trials, trial "0" had the best performance, with a mIoU of 0.796, and a loss of 0.107. This model was trained with a relatively higher learning rate of 0.001, a larger batch size of 4, and 40 epochs, which is a relatively small number of epochs. This

combination of hyperparameters (higher learning rate, larger batch size, and smaller number of epochs) was optimal for this specific model and dataset.

Table 4. 1 Results of Hyperparameter Tuning Training Trials for the DeepLabV3 Model. The training is implemented with K-fold cross-validation (k=5), utilizing the training dataset which includes a total of 1,450 images, with 900 images from Kvasir-SEG, and 550 images from CVC-ClinicDB.

Trial	Learning Rate	Batch Size	Number of Epochs	Loss	mIoU
0	**0.001**	**8**	**40**	**0.107**	**0.796**
1	0.0001	4	40	0.143	0.735
2	0.001	4	100	0.137	0.788
3	0.001	8	70	0.143	0.778

After selecting the model with the best validation performance, it was then evaluated using mean Intersection over Union (mIoU), on both seen and unseen datasets, and the findings are reported in Table 4. 2. As demonstrated, the model achieved acceptable mIoU values on the two seen datasets: CVC-ClinicDB and Kvasir, but with a noticeable drop in mIoU, compared to validation results. On the unseen datasets, mIoU drops significantly, indicating poor generalization strength.

Table 4. 2 Results of Testing DeepLabV3 Model with the Five Testing Datasets. The testing set includes 798 images from the remaining images from the five utilized datasets: the seen testing dataset includes 100 images from Kvasir-SEG and 62 images from CVC-ClinicDB, and the unseen testing dataset includes 380 images from CVC-ColonDB, 196 images from ETIS, and 60 images from EndoScene.

Dataset		mIoU
Seen Datasets	CVC-ClinicDB	0.7268
	Kvasir	0.7296
Unseen Datasets	CVC-ColonDB	0.5077
	EndoScene	0.6496
	ETIS-LaribPolypDB	0.4268

Figure 4. 8 presents DeepLabV3 results, across the five datasets (CVC-ClinicDB, Kvasir, CVC-ColonDB, EndoScene, and ETIS). Results show variable degrees of success in

predicting polyp masks. On the seen datasets (CVC-ClinicDB and Kvasir), predicted masks are reasonably accurate, where the result with the CVC-ClinicDB image was more accurate than Kvasir, while with the Kvasir sample, the model over-segmented the polyp.

DeepLabV3 shows non-optimal performance when applied to unseen datasets, where although the predicted masks come close to those observed, there are discrepancies in shape and size. This suggests that DeepLabV3 struggled to generalize effectively.

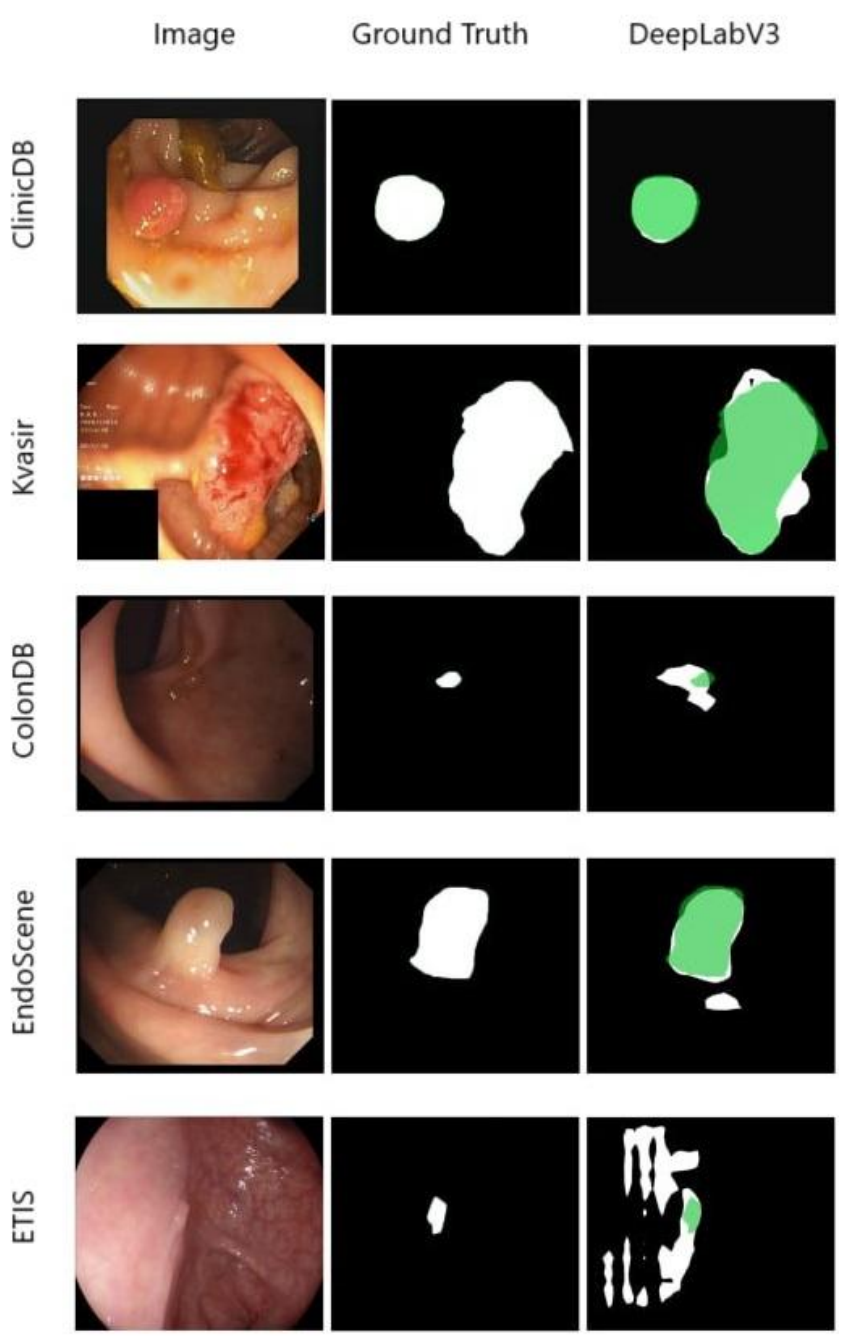

Figure 4. 8 Qualitative Results of DeepLabV3 Polyp Segmentation Model. The figure includes the original image in the first column, the ground truth mask in the second column, and the predicted mask with the U-Net model in the third column. The ground truth mask was also included in green in the third column, to provide a more precise comparison between the ground truth and predicted masks. Those results are obtained across all five testing datasets, both seen datasets (ClinicDB, Kvasir) and unseen datasets (ColonDB, EndoScene, ETIS).

4.4.2 U-Net

U-Net architecture was implemented for the image segmentation task, as a second CNN-based segmentation model. In this implementation, we used a pre-trained ResNet152 model as the encoder backbone to improve feature extraction from the input images. The same datasets with the same split were utilized, and the training pipeline followed the same structure as with the other models, including data preprocessing, augmentation techniques, and splitting the dataset to ensure a comprehensive evaluation of each model's performance in the segmentation task.

For training, we implemented k-fold cross-validation with k=5, with the combined loss function ($Loss_{combined\_BCE\_Dice}$). Online image augmentation was performed, where a random set of transformations are applied to the input data before each training epoch, to expand the dataset size and variety, to eventually increase the model's performance, robustness, and generalization ability.

RayTune was utilized For hyperparameter tuning, and Table 4. 3 outlines the results of six trials of training the model. The hyperparameters that were varied are batch size, number of epochs, and learning rate, and the evaluation metrics used to compare the different trials are $Loss_{combined\_BCE\_Dice}$ and mIoU.

Among the different trials, trial "2" had the best performance, with a mIoU of 0.794, and a loss of 0.123. This model was trained with a relatively low learning rate of 0.00013, a smaller batch size of 8, and 40 epochs, which is a relatively small number of epochs. This combination of hyperparameters (lower learning rate, smaller batch size, and smaller number of epochs) was optimal for this specific model and dataset.

Table 4. 3 Results of Hyperparameter Tuning Training Trials for the U-Net Model. The training is implemented with K-fold cross-validation (k=5), utilizing the training dataset which includes a total of 1,450 images, with 900 images from Kvasir-SEG, and 550 images from CVC-ClinicDB.

Trial	Learning Rate	Batch Size	Number of Epochs	Loss	mIoU
0	0.00056	16	100	0.134	0.736
1	0.00158	16	70	0.308	0.464
2	**0.00013**	**8**	**40**	**0.123**	**0.794**
3	0.00261	8	40	7.056	0.000
4	0.00462	8	70	7.056	0.000
5	0.00023	8	70	0.116	0.787

After selecting the model with the best validation performance, it was then evaluated using mean Intersection over Union (mIoU), on both seen and unseen datasets, and the findings are reported in Table 4. 4. As demonstrated, the model achieved acceptable mIoU values on the two seen datasets: CVC-ClinicDB and Kvasir, but with a minor drop in mIoU compared to validation results. On the unseen datasets, mIoU drops further, indicating poor to average generalization ability.

Table 4. 4 Results of Testing U-Net Model with the Five Testing Datasets. The testing set includes 798 images from the remaining images from the five utilized datasets: the seen testing dataset includes 100 images from Kvasir-SEG, 62 images from CVC-ClinicDB, and the unseen testing dataset includes 380 images from CVC-ColonDB, 196 images from ETIS-Larib and 60 images from EndoScene.

Dataset		mIoU
Seen Datasets	CVC-ClinicDB	0.7612
	Kvasir	0.7694
Unseen Datasets	CVC-ColonDB	0.5718
	EndoScene	0.7335
	ETIS-LaribPolypDB	0.4889

Figure 4. 9 displays visual results from the U-Net model across the five testing datasets: CVC-ClinicDB, Kvasir, CVC-ColonDB, EndoScene, and ETIS. On the seen datasets, the model performs relatively well. CVC-ClinicDB obtained accurate segmentation. Kvasir also

obtained a reasonably accurate prediction, with some differences between the original and predicted masks, that occur due to the large variation and complexity in Kvasir dataset images.

U-Net's results on unseen datasets vary considerably. On CVC-ColonDB, for example, its predicted mask is close to the actual mask, with some slight size variances. On EndoScene it captures most of the polyp accurately, but with some shape inconsistencies; but with ETIS, U-Net failed to segment the polyp, and produced several small regions.

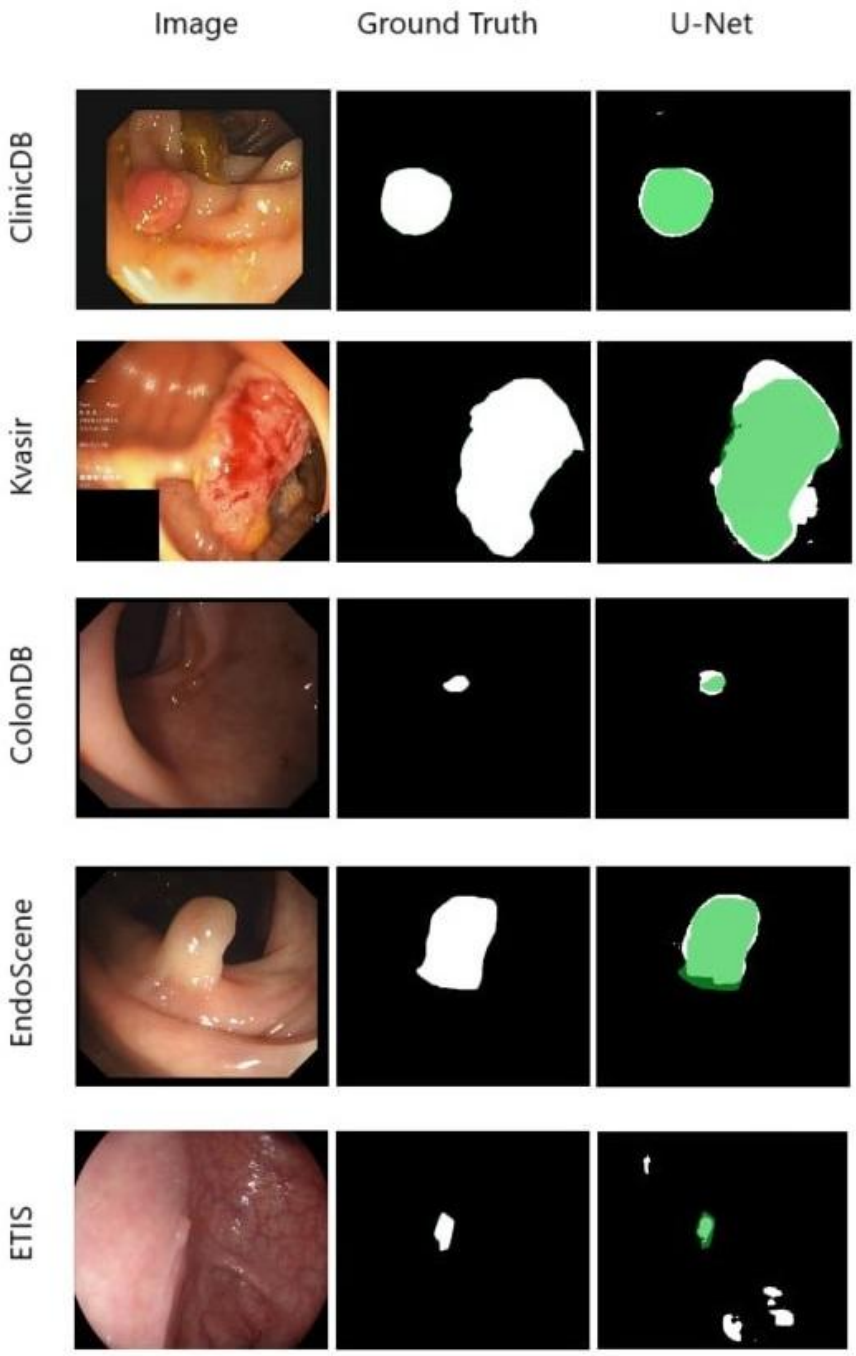

Figure 4. 9 Qualitative Results of U-Net Polyp Segmentation Model. The figure includes the original image in the first column, the ground truth mask in the second column, and the predicted mask with the U-Net model in the third column. The ground truth mask was also included in green in the third column, to provide a more precise comparison between the ground truth and predicted masks. Those results are obtained across all five testing datasets, both seen datasets (ClinicDB, Kvasir) and unseen datasets (ColonDB, EndoScene, ETIS).

4.4.3 Ensemble Model of CNN-Based Models

To leverage the capabilities of the two CNN-based models: DeepLabV3 and U-Net, a semantic segmentation ensemble model was implemented. The ensemble model is constructed by integrating the logits generated by each individual model, weighted by user-specified weights, and averaging them. This approach aims to leverage the strengths of both models, potentially providing an enhanced segmentation performance.

$$Logits_{Ensemble\_CNNs} = \frac{\left(W1 * Logits_{DeepLabV3} + W2 * Logits_{U-Net}\right)}{(W1 + W2)} \qquad (4.1)$$

$Logits_{Ensemble\_CNNs}$ is the logits generated by the ensemble model, $Logits_{DeepLabV3}$ is generated by the DeepLabV3 model, and $Logits_{U-Net}$ is generated by the U-Net model; and $W1$ and $W2$ represent user-specified weights for each model used, where we experimented and evaluated the cross-validation results with multiple weights, as outlined in Table 4. 5, to identify the model with the best performance. The results indicate that assigning the value 0.3 to $Logits_{DeepLabV3}$ and the value 0.7 to $Logits_{U-Net}$ results in designing the model with the best performance.

Table 4. 5 Results of Cross-Validation of Ensemble_CNNs Model with Multiple Weight Combinations.

Ensemble Weights	**mIoU**
($W1$=0.3, $W2$=0.7)	**0.8959**
($W1$=0.5, $W2$=0.5)	0.8833
($W1$=0.7, $W2$=0.3)	0.8720

After selecting the model with the best validation performance, it was then evaluated using mIoU, on both seen and unseen datasets, and the findings are reported in Table 4. 6. As illustrated, the model achieved relatively strong mIoU values on the two seen datasets: CVC-ClinicDB and Kvasir, but with a noticeable drop in mIoU compared to validation results, which might be interpreted as an overfitting tendency. On the unseen datasets, mIoU values indicate moderate generalization ability.

Table 4. 6 Results of Testing Ensemble_CNNs Model with the Five Testing Datasets. The testing set includes 798 images from the remaining images from the five utilized datasets: the seen testing dataset includes 100 images from Kvasir-SEG, 62 images from CVC-ClinicDB, and the unseen testing dataset includes 380 images from CVC-ColonDB, 196 images from ETIS-Larib and 60 images from EndoScene.

Dataset		**mIoU**
Seen Datasets	CVC-ClinicDB	0.8022
	Kvasir	0.8066
Unseen Datasets	CVC-ColonDB	0.6274
	EndoScene	0.7997
	ETIS-LaribPolypDB	0.5616

Figure 4. 10 provides a comparative visualization of the performance of Ensemble_CNNs on both seen and unseen datasets. On seen datasets, the predicted masks show high overlap with actual masks; this indicates the model can generalize well on familiar data. Their smooth contours closely match ground truth; particularly in Kvasir, where the model handled its complex polyps' shapes effectively.

However, it was challenging for the model to segment unseen datasets, which is made clear due to the noticeable differences between the actual and predicted masks, where the predicted masks are either under-segmenting the polyps or misaligning the mask boundaries. In ETIS, the predicted mask under-segments the polyp.

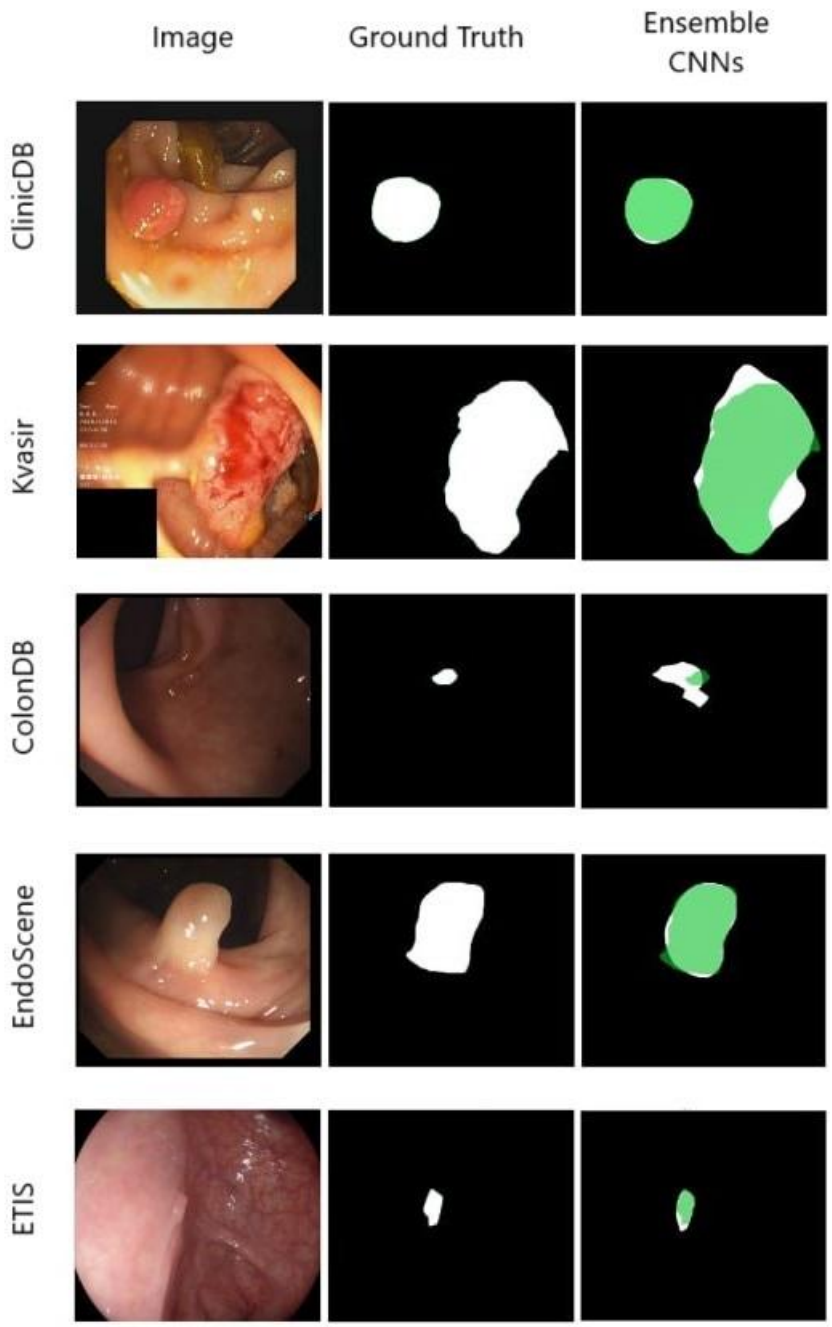

Figure 4. 10 Qualitative Results of Ensemble_CNNs Polyp Segmentation Model. The figure includes the original image in the first column, the ground truth mask in the second column, and the predicted mask with the U-Net model in the third column. The ground truth mask was also included in green in the third column, to provide a more precise comparison between the ground truth and predicted masks. Those results are obtained across all five testing datasets, both seen datasets (ClinicDB, Kvasir) and unseen datasets (ColonDB, EndoScene, ETIS). Results of Transformer-Based Models

4.5 Results of Transformer-Based Models

4.5.1 SegFormer

SegFormer was deployed as a transformer-based segmentation model. The pre-trained version of SegFormer, SegFormer-B5, was implemented to obtain superior performance, and the model's configuration settings remained as the default values.

Once the pre-trained model is loaded, it is fine-tuned with our dataset. For training, we implemented k-fold cross-validation with k=5, with the combined loss function ($Loss_{combined\_BCE\_Dice}$). Online image augmentation was performed, where a random set of transformations are applied to the input data before each training epoch, to expand the dataset size and variety, to eventually increase the model's performance, robustness, and generalization ability.

RayTune was utilized for hyperparameter tuning, and Table 4. 7 shows the results of several trials of training the model. The hyperparameters that were varied are batch size, number of epochs, and learning rate, and the evaluation metrics used to compare the different trials are $Loss_{combined\_BCE\_Dice}$ and mIoU.

Among the different trials, trial "2" had the best performance, with a mIoU of 0.8594, and a loss of 0.0874. This model was trained with a low learning rate of 0.0001, a small batch size of 4, and 100 epochs. This combination of hyperparameters (low learning rate, small batch size, and sufficient epochs) was optimal for this specific model and dataset.

Table 4. 7 Results of Hyperparameter Tuning Training Trials for the SegFormer Model. The training is implemented with K-fold cross-validation (k=5), utilizing the training dataset which includes a total of 1,450 images, with 900 images from Kvasir-SEG, and 550 images from CVC-ClinicDB.

Trial	Learning Rate	Batch Size	Number of Epochs	Loss	mIoU
0	0.001	4	70	0.1248	0.7255
1	0.0001	4	40	0.1029	0.7593
2	**0.0001**	**4**	**100**	**0.0874**	**0.8594**
3	0.001	8	40	0.872	0.8554

After selecting the model with the best validation performance, it was then evaluated using mean Intersection over Union (mIoU), on both seen and unseen datasets, and the findings are reported in Table 4. 8. As outlined, the model achieved high mIoU values on the two seen datasets: CVC-ClinicDB and Kvasir, with negligible drop in mIoU, concerning validation results. On the unseen datasets, mIoU drops as expected, but is still relatively strong performance, and is comparable with values obtained in other benchmark studies. Overall, the results are relatively strong and indicate that the model was able to learn to generalize well from the training data.

Table 4. 8 Results of Testing SegFormer Model with the Five Testing Datasets. The testing set includes 798 images from the remaining images from the five utilized datasets: the seen testing dataset includes 100 images from Kvasir-SEG, 62 images from CVC-ClinicDB, and the unseen testing dataset includes 380 images from CVC-ColonDB, 196 images from ETIS-Larib and 60 images from EndoScene.

Dataset		mIoU
Seen Datasets	CVC-ClinicDB	0.8556
	Kvasir	0.8314
Unseen Datasets	CVC-ColonDB	0.6921
	EndoScene	0.8019
	ETIS-LaribPolypDB	0.6113

Figure 4. 11 displays visual results from the SegFormer model across the five testing datasets: CVC-ClinicDB, Kvasir, CVC-ColonDB, EndoScene, and ETIS. As shown, SegFormer achieved strong performance on seen datasets, particularly CVC-ClinicDB, where its predicted polyp masks closely align with actual ones, indicating it has effectively learned to segment polyps from training data, but minor differences still exist when handling challenging images as in Kvasir. Overall, the results on seen datasets demonstrate SegFormer's ability to generalize effectively within its training dataset.

On the three unseen datasets, CVC-ColonDB, EndoScene, and ETIS, the model performed relatively well for CVC-ColonDB and EndoScene, however, ETIS proved much harder, where only small portions of polyps were captured by its predicted mask.

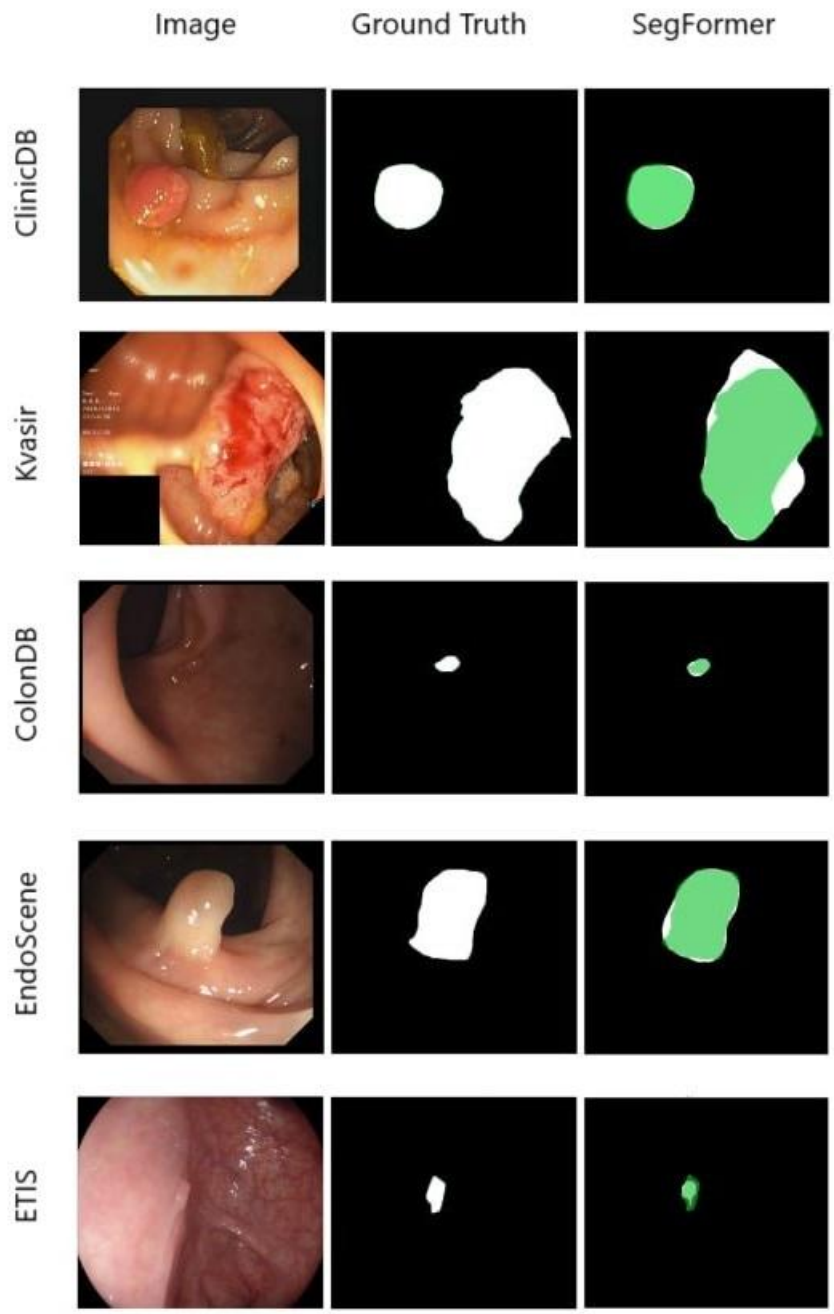

Figure 4. 11 Qualitative Results of SegFormer Polyp Segmentation Model. The figure includes the original image in the first column, the ground truth mask in the second column, and the predicted mask with the U-Net model in the third column. The ground truth mask was also included in green in the third column, to provide a more precise comparison between the ground truth and predicted masks. Those results are obtained across all five

testing datasets, both seen datasets (ClinicDB, Kvasir) and unseen datasets (ColonDB, EndoScene, ETIS).

4.5.2 Swin Transformer

Swin Transformer was deployed as another transformer-based segmentation model. For implementation, the same pipeline that was detailed earlier for modeling and dataset preparation with other models was utilized. Specifically, the smallest pre-trained model, Swin -T, was utilized as the segmentation model's backbone, since it has proven to achieve excellent results on various computer vision tasks, without compromising computational efficiency. However, UperNet was used with Swin Transformer as the backbone model. The configuration of the backbone model played a crucial role in achieving accurate segmentation results. The model was implemented with default settings and configurations.

For training, we implemented k-fold cross-validation with k=5, with the loss function $Loss_{combined\_BCE\_Dice}$. Online image augmentation was performed, where a random set of transformations are applied to the input data before each training epoch, to expand the dataset size and variety, to eventually increase the model's performance, robustness, and generalization ability.

For training hyperparameter tuning, the same methodology that was adopted with SegFormer was followed, where RayTune is utilized with the same tuning space used when tuning SegFormer. Table 4. 9 shows the results of several trials of training the model. Between trials, the hyperparameters that were varied are batch size, number of epochs, and learning rate, and the evaluation metrics used to compare the different trials are $Loss_{combined\_BCE\_Dice}$ and mIoU.

Among the different trials, trial "3" had the best performance, with a mIoU of 0.8412, and a loss of 0.0814. This model was trained with a low learning rate of 0.0001, a small batch size of 4, and 100 epochs. This combination of hyperparameters (low learning rate, small batch size, and sufficient epochs) was optimal for this specific model and dataset.

Table 4. 9 Results of Hyperparameter Tuning Training Trials for the Swin Transformer Model. The training is implemented with K-fold cross-validation (k=5), utilizing the training dataset which includes a total of 1,450 images, with 900 images from Kvasir-SEG, and 550 images from CVC-ClinicDB.

Trial	Learning Rate	Batch Size	Number of Epochs	Loss	mIoU
0	0.001	8	40	0.3853	0.2205
1	0.001	8	70	0.2893	0.3559
2	0.0001	4	40	0.1941	0.5494
3	**0.0001**	**4**	**100**	**0.0814**	**0.8412**

After selecting the model with the best validation performance, it was then evaluated using mean Intersection over Union (mIoU), on both seen and unseen datasets, and the findings are reported in Table 4. 10. As observed, the model achieved high mIoU values on the two seen datasets: CVC-ClinicDB and Kvasir, with only a slight drop in mIoU, compared to validation results. On the unseen datasets, mIoU drops, but is still within an acceptable range, and is still consistent with values obtained in benchmark-reviewed studies. Overall, the results obtained are relatively positive, and there was no major drop in mIoU values of the model, especially with the seen datasets, which indicates the model's ability to learn and generalize well from the training data.

Table 4. 10 Results of Testing the Swin Transformer Model with the Five Testing Datasets. The testing set includes 798 images from the remaining images from the five utilized datasets: the seen testing dataset includes 100 images from Kvasir-SEG, 62 images from CVC-ClinicDB, and the unseen testing dataset includes 380 images from CVC-ColonDB, 196 images from ETIS-Larib and 60 images from EndoScene.

Dataset	mIoU

Seen Datasets	CVC-ClinicDB	0.8147
	Kvasir	0.8211
Unseen Datasets	CVC-ColonDB	0.5694
	EndoScene	0.7641
	ETIS-LaribPolypDB	0.5532

Figure 4. 12 displays visual results from the Swin Transformer model across the five testing datasets: CVC-ClinicDB, Kvasir, CVC-ColonDB, EndoScene, and ETIS-LaribPolypDB.

On the seen datasets, CVC-ClinicDB and Kvasir, the model performs as expected with fairly accurate predictions. With CVC-ClinicDB, its predictions align closely with actual masks indicating it has learned to recognize polyps. However, with Kvasir, and due to variations in appearance among polyps in this dataset, it was more challenging for the model to segment it, where it over-segmented the polyp.

On the other hand, performance on the unseen datasets (CVC-ColonDB, EndoScene, and ETIS-LaribPolypDB) reflects its generalization capabilities. CVC-ColonDB and EndoScene data sets show a generally satisfactory prediction accuracy with polyp regions closely matching predicted masks generated by the model, which is especially important considering it wasn't exposed to this data during training. The predicted mask of the ETIS image does not accurately represent polyp shapes with significant differences between predicted and actual masks, suggesting the model has difficulties when applied to datasets with more irregular polyp shapes.

Overall, the model shows strong performance on seen data, and satisfactory generalization to some unseen datasets. However, additional fine-tuning might help obtain better generalization.

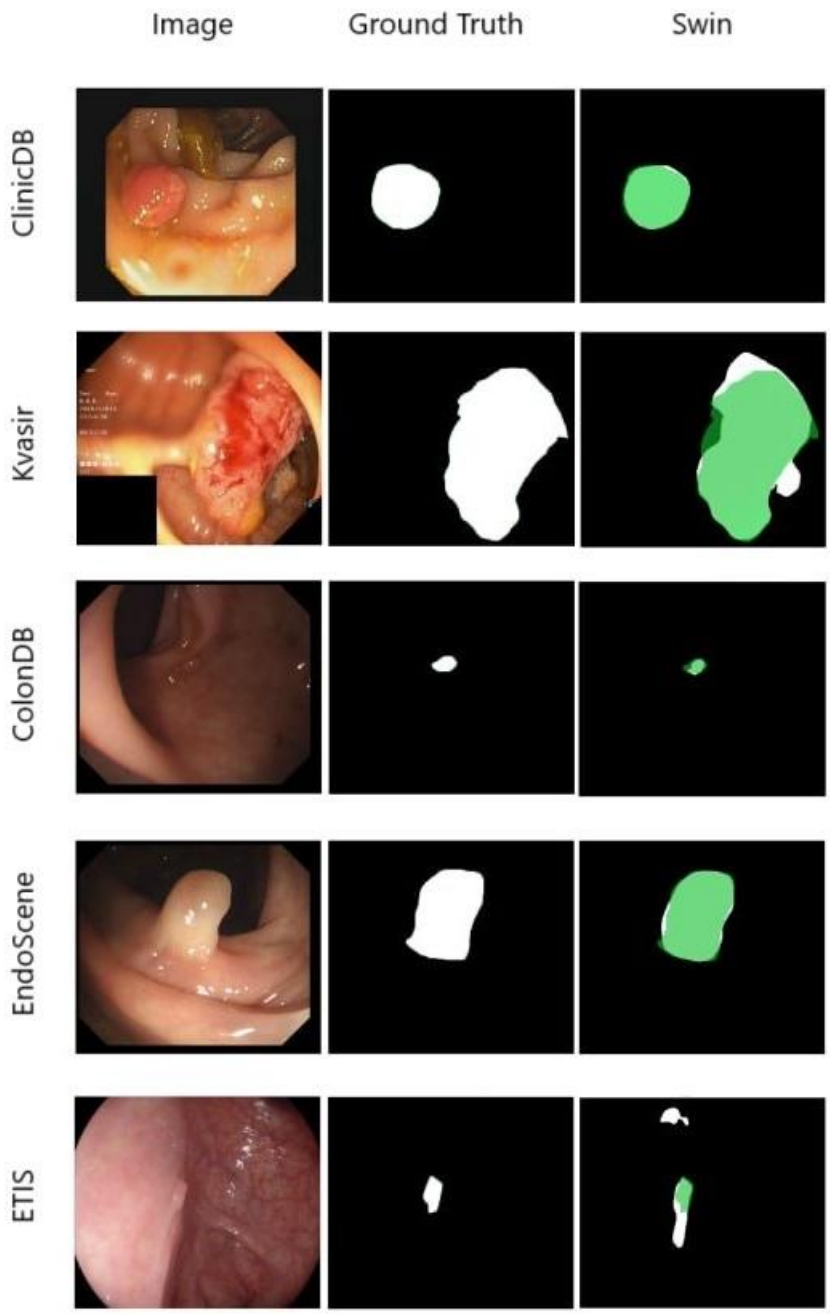

Figure 4. 12 Qualitative Results of Swin Transformer Polyp Segmentation Model. The figure includes the original image in the first column, the ground truth mask in the second column, and the predicted mask with the U-Net model in the third column. The ground truth mask was also included in green in the third column, to provide a more precise comparison between the ground truth and predicted masks. Those results are obtained across all five testing datasets, both seen datasets (ClinicDB, Kvasir) and unseen datasets (ColonDB, EndoScene, ETIS).

4.5.3 Ensemble Model of Transformer-Based Models

In this study, a semantic segmentation ensemble model was implemented to leverage the capabilities of the two state-of-the-art models: SegFormer and UperNet with the Swin Transformer backbone. The ensemble model is built by integrating the logits generated by each model, weighted by user-specified weights, and averaging them. The purpose of this approach is to exploit the complementary strengths of the two models, potentially leading to improved segmentation performance.

The weighted combination of logits from the two models can be represented mathematically as follows:

$$Logits_{Ensemble\_Transformers} = \frac{(W1 * Logits_{SegFormer} + W2 * Logits_{Swin})}{(W1 + W2)} \quad (4.1)$$

where $Logits_{Ensemble\_Transformers}$ denotes the logits generated by the ensemble model, $Logits_{SegFormer}$ and $Logits_{Swin}$ represent the logits produced by the SegFormer and Swin models, respectively, and $W1$ and $W2$ are the user-specified weights for each model, where we experimented and evaluated the cross-validation results with multiple weights, to identify the model with the best performance. The results indicate that assigning the value 0.7 to $Logits_{SegFormer}$ and the value 0.7 to $Logits_{Swin}$ results in obtaining the best performance.

Table 4. 11 Results of Cross-Validation of Ensemble_Transformers Model, with Multiple Weight Combinations.

Ensemble Weights	mIoU
($W1$=0.3, $W2$=0.7)	0.8527
($W1$=0.5, $W2$=0.5)	0.8679
($W1$=0.7, $W2$=0.3)	**0.8777**

After selecting the model with the best validation performance, it was then evaluated using mIoU, on both seen and unseen datasets, and the findings are reported in Table 4. 2. As

illustrated, the model achieved relatively strong mIoU values on the two seen datasets: CVC-ClinicDB and Kvasir, with negligible drop in mIoU compared to validation results, which indicates a robust performance and generalization, with no overfitting. On the unseen datasets, mIoU values indicate superior generalization ability.

Table 4. 12 Results of Testing Ensemble_Transformers Model with the Five Testing Datasets. The testing set includes 798 images from the remaining images from the five utilized datasets: the seen testing dataset includes 100 images from Kvasir-SEG, 62 images from CVC-ClinicDB, and the unseen testing dataset includes 380 images from CVC-ColonDB, 196 images from ETIS-Larib and 60 images from EndoScene.

Dataset		**mIoU**
Seen Datasets	CVC-ClinicDB	0.8795
	Kvasir	0.8478
Unseen Datasets	CVC-ColonDB	0.7047
	EndoScene	0.8314
	ETIS-LaribPolypDB	0.6581

Figure 4. 13 displays the segmentation results of the Ensemble_Transformers model applied to both seen and unseen datasets for polyp detection. The results on the seen datasets demonstrate a strong overlap between the predicted and actual masks, but with some over-segmented parts, indicating that the model has effectively learned the polyp features and performs well on the data it was trained on. This means that the model can generalize well to the training data, capturing the essential features of polyps with high accuracy. Similar performance was observed with the unseen dataset, with some differences between the original and predicted masks, which were oversampling for some images, and undersampling for others.

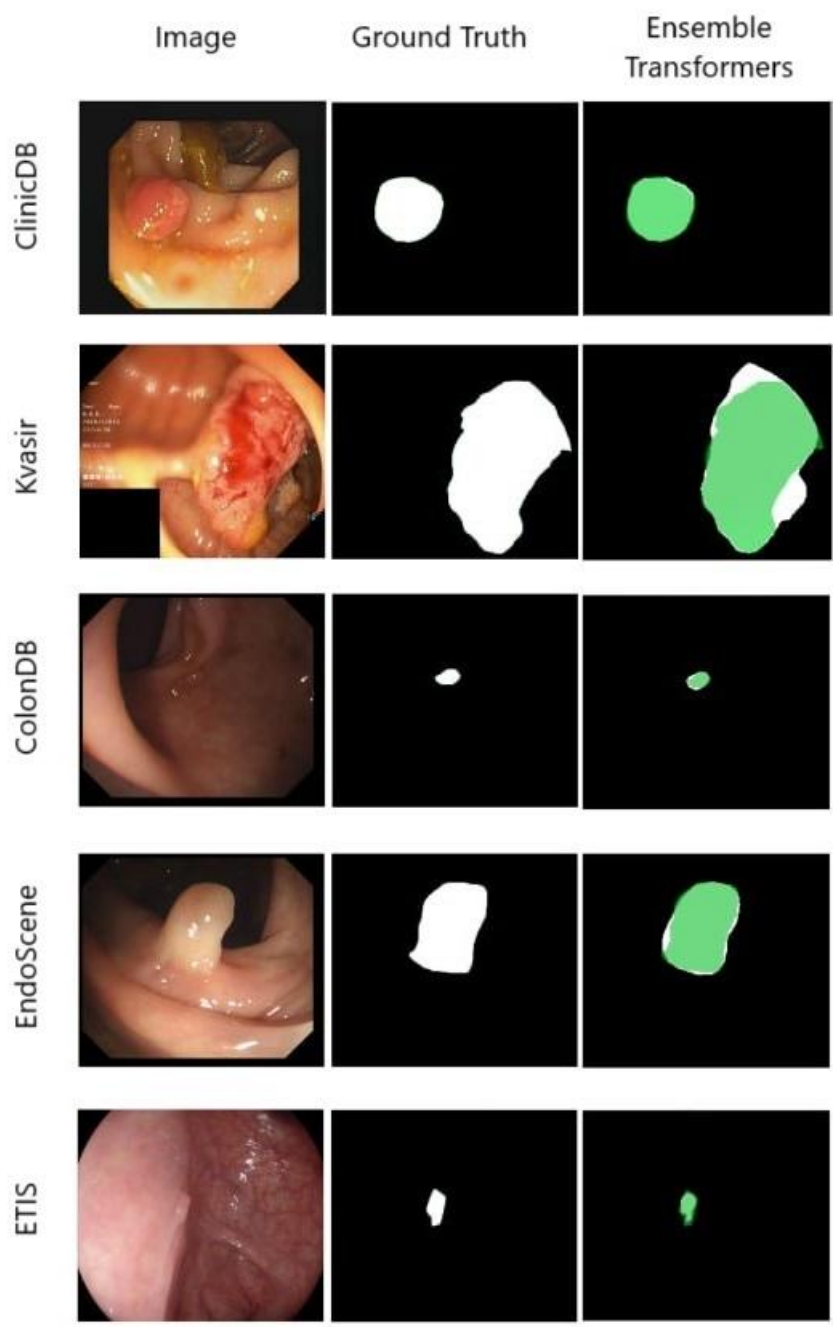

Figure 4. 13 Qualitative Results of Ensemble_Transformers Polyp Segmentation Model. The figure includes the original image in the first column, the ground truth mask in the second column, and the predicted mask with the U-Net model in the third column. The ground truth mask was also included in green in the third column, to provide a more precise comparison between the ground truth and predicted masks. Those results are obtained across all five testing datasets, both seen datasets (ClinicDB, Kvasir) and unseen datasets (ColonDB, EndoScene, ETIS).

4.6 Quantitative Results

For evaluation, a comparison is carried out for the performance of the implemented deep learning segmentation models on the five testing datasets: the testing datasets of Kvasir-SEG, and CVC-ClinicDB as seen datasets, and the ColonDB, Endoscene, and ETIS as the unseen testing datasets. Mean Intersection over Union (mIoU) served as the primary evaluation metric, where higher values indicate higher segmentation accuracy. Table 4. 13 below details each model's mIoU scores across the different datasets.

Table 4. 13 Quantitative Results of Implemented Polyp Segmentation Models (mIoU). Validation results are obtained with K-fold cross-validation (k=5), utilizing the training dataset which includes a total of 1,450 images, with 900 images from Kvasir-SEG, and 550 images from CVC-ClinicDB. The testing results are obtained for the testing dataset which includes 798 images from the remaining images from the five utilized datasets: the seen testing dataset includes 100 images from Kvasir-SEG and 62 images from CVC-ClinicDB, and the unseen testing dataset includes 380 images from CVC-ColonDB, 196 images from ETIS, and 60 images from EndoScene.

Model	**Validation**	**Testing**				
		Seen Datasets		Unseen Datasets		
		ClinicDB	Kvasir	ColonDB	EndoScene	ETIS
U-Net	0.794	0.7612	0.7694	0.5718	0.7335	0.4889
DeepLabV3	0.796	0.7268	0.7296	0.5077	0.6496	0.4268
Ensemble_CNNs	**0.8959**	0.8022	0.8066	0.6274	0.7997	0.5616
Swin-T	0.8412	0.8147	0.8211	0.5694	0.7641	0.5532
SegFormer-B5	0.8594	0.8556	0.8314	0.6921	0.8019	0.6113
Ensemble_Transformers	0.8777	**0.8795**	**0.8478**	**0.7047**	**0.8314**	**0.6581**

Referring to validation results, Ensemble_CNNs, Ensemble_Transformers, and SegFormer models achieved the highest mIoU scores. Ensemble_CNNs achieved a score of 0.8959 as the top score, followed closely by Ensemble_Transformers at 0.8777 and SegFormer with a 0.8594 validation mIoU score. Traditional U-Net and DeepLabV3 models demonstrated lower validation mIoU scores than transformer-based models at 0.794 and 0.796 respectively, suggesting low accuracy over training datasets than transformer-based models.

Testing results on the two seen datasets, ClinicDB and Kvasir, indicate that Ensemble_Transformers and SegFormer consistently outperformed other models. Ensemble_Transformers achieved the highest mIoU scores (0.8795 on ClinicDB and 0.8478 on Kvasir respectively), while SegFormer followed closely behind with 0.8556 and 0.8314 scores respectively. U-Net and DeepLabV3 both produced lower mIoU scores than traditional methods: U-Net achieved 0.7612 on ClinicDB and 0.7694 on Kvasir, while DeepLabV3 reached 0.7268 and 0.7296 for both databases respectively.

As expected, when testing results on the three unseen datasets (ColonDB, EndoScene, and ETIS), overall performance was lower across all models. Ensemble_Transformers demonstrated superior generalization ability by scoring 0.7047 on ColonDB; 0.8314 on EndoScene; and 0.6581 on ETIS respectively; SegFormer also performed well, scoring a mIoU score of 0.6921, 0.8019, and 0.6113 respectively.

U-Net and DeepLabV3 models demonstrated significantly lower performance, scoring 0.5718 on ColonDB, 0.7335 on EndoScene, and 0.4889 on ETIS respectively; DeepLabV3 performed worst overall with its mIoU score being 0.5077 across ColonDB, EndoScene, ETIS.

Transformer-based models (Ensemble_Transformers and SegFormer) proved more robust for polyp segmentation tasks in clinical applications than their CNN-based counterparts during both validation and testing. The Ensemble_Transformers model demonstrated superior generalization abilities with unseen datasets than the other models. Generally, Results indicated that transformer-based architectures have better generalization capabilities

for unseen datasets, making them suitable for use for polyp segmentation tasks in clinical applications.

Comparing the results of this study with the findings of prior research, it is evident that they are largely consistent with previous studies, particularly regarding the use of Vision Transformers for polyp segmentation. Compared to [24], our model demonstrated comparable performance on Kvasir-SEG (mIoU: 0.85 vs. 0.86) and ClinicDB (mIoU: 0.88 vs. 0.89), confirming the effectiveness of Vision Transformers on seen datasets. However, our model performed slightly worse on unseen datasets such as ColonDB (mIoU: 0.70 vs. 0.73) and ETIS (mIoU: 0.66 vs. 0.71). This performance gap underscores the challenge in generalizing to unseen data, possibly due to the effectiveness of the training strategy design. Our results on EndoScene were equal (mIoU: 0.83), suggesting that Vision Transformer consistently handles this dataset across studies.

In comparison with [23], which combines CNNs and Vision Transformers, our results indicate a slight trade-off. While our Vision Transformer model achieved slightly lower mIoU scores on Kvasir (0.85 vs. 0.87) and ClinicDB (0.88 vs. 0.89), it maintained similar performance on EndoScene (0.83). However, on unseen datasets, our model's mIoU scores for ColonDB (0.70 vs. 0.73) and ETIS (0.66 vs. 0.69) were marginally lower, further emphasizing the difficulty of generalization. Overall, our approach closely mirrors prior research, however, slight deviations demonstrate the continued need to increase generalizability and domain adaptation for polyp segmentation tasks.

Figure 4. 14 A comparison between the findings of this study, with the findings of prior research.

Study	Datasets	Task	Algorithm	Results
[24]	Training and Testing: Kvasir-SEG ClinicDB Testing: ColonDB ETIS EndoScene	Segmentation	Vision Transformer	mIoU: 0.86 (Kvasir) mIoU: 0.89 (ClinicDB) mIoU: 0.73 (ColonDB) mIoU: 0.71 (ETIS) mIoU: 0.83 (EndoScene)
[23]	Training and Testing: Kvasir ClinicDB Testing: ColonDB ETIS EndoScene	Segmentation	CNN and Vision Transformer	mIoU: 0.87 (Kvasir) mIoU: 0.89 (ClinicDB) mIoU: 0.73 (ColonDB) mIoU: 0.69 (ETIS) mIoU: 0.84 (EndoScene)
Our work	Training and Testing: Kvasir-SEG ClinicDB Testing: ColonDB ETIS EndoScene	Segmentation	Vision Transformer	mIoU: 0.85 (Kvasir) mIoU: 0.88 (ClinicDB) mIoU: 0.70 (ColonDB) mIoU: 0.66 (ETIS) mIoU: 0.83 (EndoScene)

4.7 Qualitative Results

To intuitively demonstrate the effectiveness of the models as semantic segmentation models, an analysis of the qualitative results of all implemented models, presented in Figure 4. 15, was carried out.

Most models performed well on seen datasets, producing accurate predictions that closely mirrored ground truth data. With ClinicDB, the masks produced by all models provided similar polyp shapes, which closely mirrored ground truth masks. While with Kvasir, the

results were more varied, CNN-based models (DeepLabV3 and U-Net) showed signs of over-segmentation while other models provided more accurate predictions that aligned more closely with ground truth data.

On the unseen datasets (ColonDB, EndoScene, and ETIS), and as expected, models generally struggled more than the seen datasets. DeepLabV3 and U-Net displayed significant under-segmentation while all transformer-based models (SegFormer, Swin Transformer, and their ensemble) delivered relatively better results.

With ColonDB, the masks predicted by CNN-based models are either incomplete or severely fragmented compared to the ground truth, while the results of transformer models were better.

With EndoScene, the segmentation results are mostly reasonable and consistent across models. However, CNN-based models showed some under-segmentation, while transformer models are more precise, and closer to the ground truth. ETIS, which is the most challenging dataset, had predictions that were either heavily fragmented or significantly smaller than the ground truth, especially with CNN-based models. The transformers ensemble model and SegFormer model had the most accurate predictions while still exhibiting some under-segmentation.

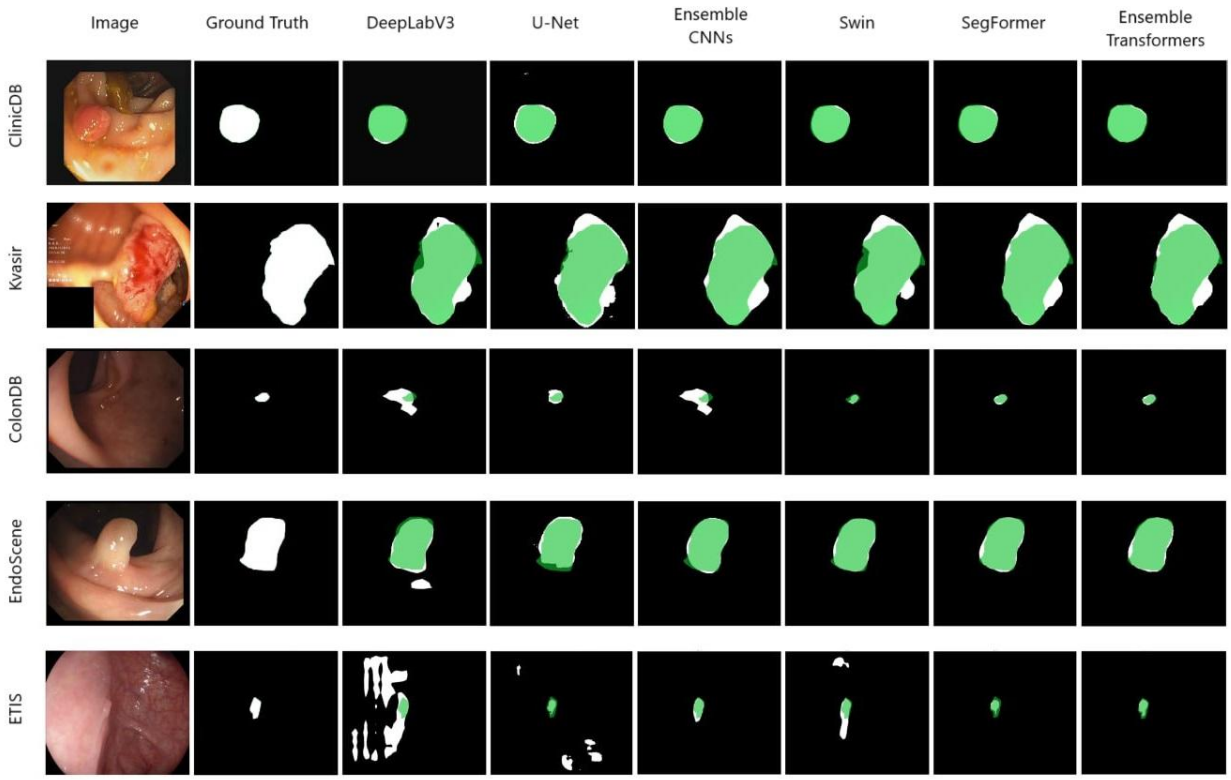

Figure 4. 15 Qualitative Results of Implemented Polyp Segmentation Models. The figure includes the original image in the first column, the ground truth mask in the second column, and the predicted mask with each model in the later columns. The ground truth mask was also included in green in the prediction columns, to provide a more precise comparison between the ground truth and predicted masks. Those results are obtained across all five testing datasets, both seen datasets (ClinicDB, Kvasir) and unseen datasets (ColonDB, EndoScene, ETIS).

4.8 Summary

This chapter explored the use of deep learning models for polyp segmentation in medical images. Experiments with data augmentation techniques were undertaken, and the performance of CNN-based models, and transformer-based models was evaluated and compared. Results illustrated both online data augmenting techniques as well as model selection considerations as key factors to accurate polyp segmentation.

When testing all models on both seen and unseen datasets, transformer-based models showed superior performance. Ensemble_Transformers outperformed all other models on the seen

datasets, scoring a mIoU of 0.8795 on ClinicDB, and 0.8478 on Kvasir, as well as with the unseen datasets, scoring a mIoU of 0.7047 on ColonDB, 0.8314 on EndoScene, and 0.6581 on ETIS. SegFormer followed closely, demonstrating strong performance as well. On the other hand, U-Net and DeepLabV3 showed significantly lower performance, particularly on unseen datasets, indicating that transformer-based models generalize better and are more suitable for polyp segmentation in clinical applications.

Lastly, the qualitative results of the models were analyzed to intuitively demonstrate the effectiveness of the models as semantic segmentation models.

On seen datasets, most models including traditional CNNs and transformers produced accurate segmentation masks that closely aligned with ground truth with ClinicDB data, while results varied for Kvasir, where polyps were over-segmented by CNN-based models, while transformer-based predictions provided more accurate predictions.

On the unseen datasets, models generally struggled, particularly the CNN-based models. On ColonDB, CNN-based models produced fragmented and incomplete masks, while transformer-based models outshone them. On EndoScene, transformer models were more consistent and accurate compared to CNNs which displayed under-segmentation. ETIS proved particular predictions CNNs produced fragmented or undersized masks while Ensemble_Transformers and SegFormer models generated accurate predictions despite showing some under-segmentation.

Vision transformer-based models outperformed CNN-based models in our polyp segmentation task due to their adopted global attention mechanism, which captures long-range dependencies across an entire image. Due to the high variance of polyps in terms of

size, shape, and texture, it is crucial to be able to distinguish the polyp from its surrounding tissues, and vision transformers accomplish this by modeling relationships between distant image regions, unlike CNNs with their local receptive fields that often struggle with such variation.

Additionally, vision transformers managed to maintain more precise details due to their multi-head self-attention mechanism, which prevents the loss of spatial resolution, that CNNs experience due to their down-sampling operations. This enables transformers to handle different feature scales across images for better boundary detection and segmentation, which is particularly important when dealing with small and irregular polyps.

Chapter Five: Conclusion and Future Work

5.1 Conclusion

In conclusion, Statistics showing cancer as the second leading cause of death worldwide, and colorectal cancer being second in terms of cancer deaths, shows how urgently proactive actions need to be taken. Computer vision, as an automated diagnostic tool replacing time-consuming screening of colonoscopy videos by experts, has become an essential requirement to help address these major concerns.

Vision Transformer models have outperformed CNNs in the latest studies, and therefore they were the core of the Deep Learning image segmentation pipelines designed in this work. This thesis investigated the potential of Vision Transformer-based deep learning models for automated polyp segmentation of colonoscopy images, to address the limitations of traditional human-based polyp detection, including subjectivity, fatigue, and accordingly, miss rates. U-Net and DeepLab with the ResNet backbone, along with their ensemble model, were implemented as CNN-based baseline models. On the other hand, we implemented two ViT models, SegFormer and UperNet with the Swin Transformer backbone, as well as an ensemble model of the two, and compared the results obtained by all models.

During validation, the Ensemble_CNNs model achieved the highest mIoU of 0.8959, while the three Transformer-based models (Ensemble_Transformers, SegFormer, and Swin transformer) achieved the second, third, and fourth highest mIoU scores of 0.8777, 0.8594, and 0.8412, respectively. Testing on seen datasets further demonstrated the superiority of Vision Transformers, with Ensemble_Transformers achieving the highest mIoU scores of 0.8795 on ClinicDB and 0.8478 on Kvasir. Among all other models, Transformer-based

models achieved the highest performance. Additionally, on unseen datasets, all Transformer-based models displayed superior generalization ability, where Ensemble_Transformers achieved the highest mIoU scores of 0.7047 on ColonDB, 0.8314 on EndoScene, and 0.6581 on ETIS, significantly outperforming CNN-based models like U-Net (0.5718, 0.7335, 0.4889) and DeepLabV3 (0.5077, 0.7335, 0.6581).

These findings demonstrate that Vision Transformer-based models are more robust and generalize better in polyp segmentation tasks in both seen and unseen datasets, making them suitable for clinical applications.

This is primarily due to using global attention mechanisms by vision transformers, that capture long-range dependencies, which is essential to distinguish polyps of various sizes, shapes, and textures from surrounding tissues. Furthermore, transformers' multi-head self-attention preserves spatial details across feature scales, providing more accurate boundary detection and segmentation, especially of irregularly shaped polyps.

By exploring an underexplored area of medical image analysis, this research advances automated diagnostic tools while contributing to the early detection and prevention of colorectal cancer.

5.2 Limitations and Future Work

This study has limitations that offer opportunities for future research directions, which include the following:

- Limited dataset size and inherent variability: Deep learning models rely heavily on training data of sufficient quantity and quality. While five publicly available datasets were used for training and testing purposes, their inherent variability may limit performance due to factors like equipment differences, colonoscopy conditions, or setting variations. This can be tackled by expanding to a larger dataset with wider polyp variations, as well as adopting weakly-supervised, semi-supervised, and self-supervised learning.
    - Expanding to a larger set with wider polyp variations could potentially improve model generalizability and performance when dealing with real-world colonoscopy data.
    - Collecting large and variant medical datasets is challenging due to privacy restrictions and the time-consuming annotation process. To overcome these data limitations, incorporating weakly, semi and self-supervised learning methods in the implementation pipeline can be explored to mitigate the impact of this limitation.
- Real-world data integration: This work utilizes publicly available datasets that might not cover all the complexities and variations present in real-world colonoscopy videos. Future work could involve collaboration with hospitals in order to incorporate real-world data for training and validation, designing models more adaptable to real-life situations.
- Computational cost: Training ViT models are often computationally intensive compared to CNNs. Exploring techniques for efficient training and inference may prove invaluable when working with larger and more computationally demanding transformers. Additionally, this also facilitates real-world deployment.

- Clinical evaluation: While the quantitative and qualitative results of this work are promising, an important next step would be conducting clinical evaluations with actual gastroenterologists to test the performance of the designed models in a real-world setting. This would involve integrating them into existing colonoscopy workflows and gathering feedback from medical professionals.

References

C. Mattiuzzi and G. Lippi, “Current Cancer Epidemiology,” J Epidemiol Glob Health, vol. 9, no. 4, p. 217, 2019, doi: 10.2991/jegh.k.191008.001.

“World Health Organization (WHO), ‘Colorectal Cancer.’” Accessed: Jul. 17, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/colorectal-cancer

A. M. Leufkens, M. G. Van Oijen, F. P. Vleggaar, and P. D. Siersema, “1178 Factors Affecting Miss Rate of Polyps During Colonoscopy: Results From a Prospective, Multicenter Back-to-Back Colonoscopy Study,” Gastrointest Endosc, vol. 73, no. 4, p. AB165, Apr. 2011, doi: 10.1016/j.gie.2011.03.157.

T. Sawicki, M. Ruszkowska, A. Danielewicz, E. Niedźwiedzka, T. Arłukowicz, and K. E. Przybyłowicz, “A Review of Colorectal Cancer in Terms of Epidemiology, Risk Factors, Development, Symptoms and Diagnosis,” Cancers (Basel), vol. 13, no. 9, p. 2025, Apr. 2021, doi: 10.3390/cancers13092025.

A. Richardson, J. Hayes, C. Frampton, and J. Potter, “Modifiable lifestyle factors that could reduce the incidence of colorectal cancer in New Zealand.,” N Z Med J, vol. 129, no. 1447, pp. 13–20, Dec. 2016.

S. H. Jafri and G. Mills, “Lifestyle modification in colorectal cancer patients: an integrative oncology approach.,” Future Oncol, vol. 9, no. 2, pp. 207–18, Feb. 2013, doi: 10.2217/fon.12.184.

D. Jha et al., “Real-Time Polyp Detection, Localization and Segmentation in Colonoscopy Using Deep Learning,” IEEE Access, vol. 9, pp. 40496–40510, 2021, doi: 10.1109/ACCESS.2021.3063716.

Y. Tian et al., “Detecting, Localising and Classifying Polyps from Colonoscopy Videos using Deep Learning,” Jan. 2021.

C. M. Stauffer and C. Pfeifer, Colonoscopy. 2024.

“University of Michigan Health, ‘Colon and Rectal Polyps.’” Accessed: Jul. 17, 2024. [Online]. Available: https://www.uofmhealth.org/conditions-treatments/digestive-and-liver-health/colon-and-rectal-polyps

M. Gschwantler et al., “High-grade dysplasia and invasive carcinoma in colorectal adenomas: a multivariate analysis of the impact of adenoma and patient characteristics,” Eur J Gastroenterol Hepatol, vol. 14, no. 2, pp. 183–188, Feb. 2002, doi: 10.1097/00042737-200202000-00013.

“American Cancer Society, ‘Colorectal Cancer Early Detection, Diagnosis, and Staging - Detection and Diagnosis.’” Accessed: Jul. 17, 2024. [Online]. Available: https://www.cancer.org/cancer/types/colon-rectal-cancer/detection-diagnosis-staging/detection.html

A. M. Leufkens, M. G. Van Oijen, F. P. Vleggaar, and P. D. Siersema, “1178 Factors Affecting Miss Rate of Polyps During Colonoscopy: Results From a Prospective, Multicenter Back-to-Back Colonoscopy Study,” Gastrointest Endosc, vol. 73, no. 4, Supplement, p. AB165, 2011, doi: https://doi.org/10.1016/j.gie.2011.03.157.

N. Tajbakhsh, S. R. Gurudu, and J. Liang, “Automated Polyp Detection in Colonoscopy Videos Using Shape and Context Information,” IEEE Trans Med Imaging, vol. 35, no. 2, pp. 630–644, Feb. 2016, doi: 10.1109/TMI.2015.2487997.

M. Arnold, M. S. Sierra, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, “Global patterns and trends in colorectal cancer incidence and mortality.,” Gut, vol. 66, no. 4, pp. 683–691, Apr. 2017, doi: 10.1136/gutjnl-2015-310912.

J. C. van Rijn, J. B. Reitsma, J. Stoker, P. M. Bossuyt, S. J. van Deventer, and E. Dekker, "Polyp miss rate determined by tandem colonoscopy: a systematic review.," Am J Gastroenterol, vol. 101, no. 2, pp. 343–50, Feb. 2006, doi: 10.1111/j.1572-0241.2006.00390.x.

H. Seo et al., "Machine learning techniques for biomedical image segmentation: An overview of technical aspects and introduction to state-of-art applications.," Med Phys, vol. 47, no. 5, pp. e148–e167, Jun. 2020, doi: 10.1002/mp.13649.

Y. Shin, H. A. Qadir, L. Aabakken, J. Bergsland, and I. Balasingham, "Automatic colon polyp detection using region-based deep CNN and post learning approaches," IEEE Access, vol. 6, pp. 40950–40962, Jul. 2018, doi: 10.1109/ACCESS.2018.2856402.

K. Pogorelov et al., "Deep Learning and Hand-crafted Feature Based Approaches for Polyp Detection in Medical Videos."

Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," Mar. 2021, [Online]. Available: http://arxiv.org/abs/2103.14030

E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers," May 2021, [Online]. Available: http://arxiv.org/abs/2105.15203

M. Caron et al., "Emerging Properties in Self-Supervised Vision Transformers," Apr. 2021, [Online]. Available: http://arxiv.org/abs/2104.14294

K. B. Park and J. Y. Lee, "SwinE-Net: hybrid deep learning approach to novel polyp segmentation using convolutional neural network and Swin Transformer," J Comput Des Eng, vol. 9, no. 2, pp. 616–632, Apr. 2022, doi: 10.1093/jcde/qwac018.

B. Dong, W. Wang, D.-P. Fan, J. Li, H. Fu, and L. Shao, “Polyp-PVT: Polyp Segmentation with Pyramid Vision Transformers,” Aug. 2021, [Online]. Available: http://arxiv.org/abs/2108.06932

A. Dosovitskiy et al., “AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE.” [Online]. Available: https://github.com/

Q. Angermann et al., Towards Real-Time Polyp Detection in Colonoscopy Videos: Adapting Still Frame-Based Methodologies for Video Sequences Analysis. 2017. doi: 10.1007/978-3-319-67543-5_3.

Y. Shin, H. A. Qadir, L. Aabakken, J. Bergsland, and I. Balasingham, “Automatic Colon Polyp Detection Using Region Based Deep CNN and Post Learning Approaches,” IEEE Access, vol. 6, pp. 40950–40962, 2018, doi: 10.1109/ACCESS.2018.2856402.

L. Alzubaidi et al., “Review of deep learning: concepts, CNN architectures, challenges, applications, future directions,” J Big Data, vol. 8, no. 1, p. 53, Mar. 2021, doi: 10.1186/s40537-021-00444-8.

S. R. Dubey, S. K. Singh, and B. B. Chaudhuri, “Activation Functions in Deep Learning: A Comprehensive Survey and Benchmark,” Sep. 2021.

H. KOYUNCU, “Loss Function Selection in NN based Classifiers: Try-outs with a Novel Method,” in 2020 12th International Conference on Electronics, Computers and Artificial Intelligence (ECAI), IEEE, Jun. 2020, pp. 1–6. doi: 10.1109/ECAI50035.2020.9223208.

A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, in NIPS’12. Red Hook, NY, USA: Curran Associates Inc., 2012, pp. 1097–1105.

K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” Sep. 2014.

K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” Dec. 2015.

M. Tan and Q. V. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” May 2019.

J. Long, E. Shelhamer, and T. Darrell, “Fully Convolutional Networks for Semantic Segmentation,” Nov. 2014, [Online]. Available: http://arxiv.org/abs/1411.4038

L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs,” Dec. 2014, [Online]. Available: http://arxiv.org/abs/1412.7062

L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs,” Jun. 2016, [Online]. Available: http://arxiv.org/abs/1606.00915

L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, “Rethinking Atrous Convolution for Semantic Image Segmentation,” Jun. 2017, [Online]. Available: http://arxiv.org/abs/1706.05587

L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation,” Feb. 2018, [Online]. Available: http://arxiv.org/abs/1802.02611

F. Chollet, “Xception: Deep Learning with Depthwise Separable Convolutions,” Oct. 2016, [Online]. Available: http://arxiv.org/abs/1610.02357

A. Vaswani et al., “Attention Is All You Need,” Jun. 2017.

N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-End Object Detection with Transformers,” May 2020.

X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local Neural Networks,” Nov. 2017.

H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, “Training data-efficient image transformers & distillation through attention,” Dec. 2020.

Z. Chen, J. Zou, and X. Wang, “Semantic Segmentation on VSPW Dataset through Aggregation of Transformer Models,” Sep. 2021, [Online]. Available: http://arxiv.org/abs/2109.01316

G. Leung, J. Gao, X. Zeng, and S. Fidler, “Improving Semantic Segmentation in Transformers using Hierarchical Inter-Level Attention,” Jul. 2022, [Online]. Available: http://arxiv.org/abs/2207.02126

I. J. Goodfellow et al., “Generative Adversarial Networks,” Jun. 2014, [Online]. Available: http://arxiv.org/abs/1406.2661

J. Y. Lee et al., “Real-time detection of colon polyps during colonoscopy using deep learning: systematic validation with four independent datasets.,” Sci Rep, vol. 10, no. 1, p. 8379, May 2020, doi: 10.1038/s41598-020-65387-1.

J. Redmon and A. Farhadi, “YOLO9000: Better, Faster, Stronger,” Dec. 2016, [Online]. Available: http://arxiv.org/abs/1612.08242

O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” May 2015.

D. Jha et al., “Kvasir-SEG: A Segmented Polyp Dataset,” Nov. 2019, [Online]. Available: http://arxiv.org/abs/1911.07069

J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, “WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency

maps from physicians," Computerized Medical Imaging and Graphics, vol. 43, pp. 99–111, 2015, doi: https://doi.org/10.1016/j.compmedimag.2015.02.007.

N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automated Polyp Detection in Colonoscopy Videos Using Shape and Context Information," IEEE Trans Med Imaging, vol. 35, no. 2, pp. 630–644, Feb. 2016, doi: 10.1109/TMI.2015.2487997.

D. Vázquez et al., "A Benchmark for Endoluminal Scene Segmentation of Colonoscopy Images," Dec. 2016, [Online]. Available: http://arxiv.org/abs/1612.00799

J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, "Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer.," Int J Comput Assist Radiol Surg, vol. 9, no. 2, pp. 283–93, Mar. 2014, doi: 10.1007/s11548-013-0926-3.

A. Nogueira-Rodríguez et al., "Deep Neural Networks approaches for detecting and classifying colorectal polyps," Neurocomputing, vol. 423, pp. 721–734, Jan. 2021, doi: 10.1016/j.neucom.2020.02.123.

A. Nogueira-Rodríguez et al., "Real-time polyp detection model using convolutional neural networks," Neural Comput Appl, vol. 34, no. 13, pp. 10375–10396, Jul. 2022, doi: 10.1007/s00521-021-06496-4.

J. Ansel et al., "PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation," in International Conference on Architectural Support for Programming Languages and Operating Systems - ASPLOS, Association for Computing Machinery, Apr. 2024, pp. 929–947. doi: 10.1145/3620665.3640366.

"Albumentations Documentation." Accessed: Jul. 21, 2024. [Online]. Available: https://albumentations.ai/docs/

P. Moritz et al., Ray: A Distributed Framework for Emerging AI Applications Ray: A Distributed Framework for Emerging AI Applications. 2018. [Online]. Available: https://www.usenix.org/conference/osdi18/presentation/nishihara

"NumPy." Accessed: Jul. 21, 2024. [Online]. Available: https://numpy.org/

F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Jan. 2012, [Online]. Available: http://arxiv.org/abs/1201.0490

W. Mckinney, "Data Structures for Statistical Computing in Python," Proceedings of the 9th Python in Science Conference, Jan. 2010.

M. Waskom, "seaborn: statistical data visualization," J Open Source Softw, vol. 6, no. 60, p. 3021, Apr. 2021, doi: 10.21105/joss.03021.

D.-P. Fan et al., "PraNet: Parallel Reverse Attention Network for Polyp Segmentation," Jun. 2020, [Online]. Available: http://arxiv.org/abs/2006.11392

Y. Zhang, H. Liu, and Q. Hu, "TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation," Feb. 2021, [Online]. Available: http://arxiv.org/abs/2102.08005

A. Mao, M. Mohri, and Y. Zhong, "Cross-Entropy Loss Functions: Theoretical Analysis and Applications," Apr. 2023, [Online]. Available: http://arxiv.org/abs/2304.07288

C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, "Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations.," Deep learning in medical image analysis and multimodal learning for clinical decision support : Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, held in conjunction with MICCAI 2017 Quebec City, QC,..., vol. 2017, pp. 240–248, 2017, doi: 10.1007/978-3-319-67558-9_28.

Z. Wang et al., “Revisiting Evaluation Metrics for Semantic Segmentation: Optimization and Evaluation of Fine-grained Intersection over Union,” Oct. 2023, [Online]. Available: http://arxiv.org/abs/2310.19252

الكشف عن الأورام باستخدام محوّلات الرؤية

أسماء أحمد أمين علاونة

د. محمد يعقوب

د. مجدي عودة

أ.د. حازم الحياري

ملخص

عالمياً، يعد السرطان السبب الثاني للوفاة، ومن بين إجمالي وفيات السرطان، يعد سرطان القولون ثاني أكثر الأسباب شيوعاً. قد تتحول الأورام التي تتشكل بسبب نمو غير طبيعي للخلايا على جدران القولون، إلى سرطان القولون مع مرور الوقت، وبالتالي فإن الكشف المبكر أمر بالغ الأهمية لاتخاذ إجراءات لمنع تلك الزوائد من التحول إلى سرطان، وبحال تحولها لسرطان، فإن الكشف المبكر يمكن من إجراء تدخل طبي في المراحل المبكرة من المرض. عادةً ما يتم اكتشاف الأورام بواسطة أطباء متخصصين يقومون بمراجعة وتحليل مقاطع الفيديو المسجلة خلال إجراء تنظير القولون، ونظراً لصعوبة هذه العملية واستغراقها لوقت طويل، فقد ظهرت الحاجة إلى نماذج الرؤية الحاسوبية كأداة للتشخيص الطبي. في هذا العمل، تم استكشاف تطبيق نماذج محولات الرؤية كنهج للتعرف على الأورام في صور تنظير القولون، مع مقارنة أدائها بنماذج الشبكات العصبية الاصطناعية.

شملت منهجية التنفيذ خطوات متعددة، بدءا بإعداد البيانات (صور الأورام) مثل تقسيمها وتعديلها بهدف تعزيزها وزيادة تنوعها، تلتها مراحل تدريب وضبط وتقييم نماذج التعلم العميق، حيث تم استخدام خمسة من مجموعات الصور الطبية لأورام القولون المتاحة للعامة (CVC-ClinicDB, Kvasir, CVC-ColonDB, EndoScene, ETIS) كخط أساس يمثل مجموعة

النماذج المعتمدة على الشبكات العصبية الاصطناعية، فقد تم تصميم النماذج DeepLabV3 وU-Net، بالإضافة لنموذج تجميعي يسخر أداء كل من هذين النموذجين. بالإضافة لذلك، وباعتباره محور هذا العمل، فقد تم تصميم النماذج التي تعتمد على محولات الرؤية SegFormer وSwin Transformer، بالإضافة لنموذج تجميعي يسخر أداء كل من النموذجين السابقين، لتمثل هذه النماذج مجموعة النماذج المعتمدة على محولات الرؤية. في النهاية، تم تقييم أداء جميع هذه النماذج باستخدام مجموعة بيانات جديدة لم تشملها مرحلة التدريب والضبط، لتقييم أداءها المستقبلي في التطبيقات السريرية الواقعية. أظهرت نتائج تقييم أداء النماذج خلال مرحلة التدريب أن نموذج Ensemble_CNNs حقق أعلى معدل mIoU بلغ 0.8959، بينما حققت النماذج الثلاثة القائمة على محولات الرؤية Ensemble_Transformers, SegFormer, Swin Transformer ثاني وثالث ورابع أعلى معدلات mIoU بلغت 0.8777 و0.8594 و0.8412 على التوالي. وعند اختبار أداء جميع النماذج باستخدام بيانات الاختبار، أظهرت محولات الرؤية تفوقاً واضحاً، حيث حقق نموذج Ensemble_Transformers أعلى قيم mIoU بلغت 0.8795 على ClinicDB و0.8478 على Kvasir، كما أظهرت النماذج الأخرى المعتمدة على محولات الرؤية أعلى أداء من بين جميع النماذج. أخيراً، وعند اختبار أداء النماذج باستخدام بيانات اختبار جديدة التي لم يتم تدريب النماذج على جزء منها، أثبتت جميع النماذج التي تعتمد على محولات الرؤية أداءها الفائق عند تحليل مجموعات صور جديدة تحاكي التطبيقات السريرية الواقعية، حيث حقق Ensemble_Transformers أعلى معدلات mIoU بلغت 0.7047 على ColonDB، و0.8314 على EndoScene، و0.6581 على ETIS، وبذلك تفوقت بشكل كبير على النماذج التي تعتمد على الشبكات العصبية مثل نموذج U-Net الذي حقق 0.5718 mIoU على ColonDB، و 0.7335 على EndoScene، و 0.4889 على ETIS، ونموذج DeepLabV3 الذي حقق0.5077 mIoU على ColonDB، و 0.6496 على EndoScene ، و 0.4268 على ETIS .

عند تقييم ومقارنة أداء النماذج المختلفة، تفوقت مجموعة النماذج المعتمدة على محولات الرؤية على مجموعة النماذج المعتمدة على الشبكات العصبية الاصطناعية، لاكتشاف الأورام ضمن الصور من مجموعات البيانات المختلفة، ومن ضمنها مجموعات الصور الجديدة التي تحاكي بيانات التطبيقات السريرية الواقعية. من خلال هذه الرسالة، تم استكشاف مجال بحث جديد في مجالات تحليل الصور الطبية، مما أسهم في تطوير أدوات تشخيصية آلية تساهم في الكشف المبكر والوقاية من سرطان القولون.

الكلمات المفتاحية: ذكاء اصطناعي، تصوير طبي، أورام.