



# **Gastrointestinal Polyp Segmentation**

Use Encoder-Decoder Networks for  
Gastrointestinal Polyp Segmentation

Mohammad Zamani  
Student No. 610399135

University of Tehran- CS Department

# Contents

<b>1</b>	<b>Abstract</b>	<b>2</b>
<b>2</b>	<b>Introduction</b>	<b>2</b>
<b>3</b>	<b>Methodology</b>	<b>3</b>
3.1	Methods . . . . .	3
3.2	Dataset . . . . .	3
3.3	Encoders . . . . .	3
3.3.1	MobileNet V2 . . . . .	4
3.3.2	Resnet50 . . . . .	5
3.4	Decoders . . . . .	5
3.4.1	U-Net . . . . .	5
3.4.2	U-Net++ . . . . .	6
3.4.3	PSPNet . . . . .	6
3.4.4	DeepLabV3Plus . . . . .	7
3.5	Transfer learning . . . . .	7
<b>4</b>	<b>Results</b>	<b>8</b>
4.1	Metrics . . . . .	8
4.1.1	Dice Score . . . . .	8
4.1.2	Dice Loss . . . . .	9
4.1.3	IOU Score . . . . .	9
4.2	Training Results . . . . .	10
4.3	Test Results . . . . .	13
<b>5</b>	<b>Conclusion</b>	<b>15</b>
	<b>References</b>	<b>15</b>

# 1 Abstract

Detecting abnormal tissues can be overlooked during body screening procedures including endoscopy, bronchoscopy, and colonoscopy. Colonoscopy is a routine screening procedure that can examine inside of the large intestine. However, observants might not be able to detect anomalies at initial phase. Therefore, a precise method is needed to detect the abnormalities. In this project we want to use some pretrained different convolutional neural networks with different encoders and decoders to segment polyps in gastrointestinal tract images, we will test and evaluate some models like UNET, UNET++, DeepLab and PSPNet with different encoders and finally combine them to reach state-of-the-art performance results.

# 2 Introduction

Colorectal cancer is one of the most common malignancies of the digestive tract and a leading cause of cancer-related death in both sexes. Polyps are abnormal growths of tissue from the mucous membrane. Colorectal polyps are usually benign, while some might be precancerous or even cancerous over a period of 5 to 15 years. Colorectal polyps can become lethal in their later stages. Therefore, it is critical to detect and remove them at an earlier stage.

Developments in computer vision and artificial intelligence algorithms have influenced various fields of science and industry. They have simplified the most complicated problems encountered in a wide variety of professions, including engineering, art, and medicine. The trend shows an increasing demand for the use of automation for a variety of tasks in real operations.

The idea behind image segmentation is to assign the same label to pixels that have the same specific characteristics. Image segmentation and object detection techniques can divide an image into meaningful parts that are easier to analyze and interpret. Image segmentation techniques can assist in detecting abnormalities in medical images. These techniques can aid pathologists in many ways, such as monitoring and improving diagnostic ability. In this paper, we have experimented with eight different architectures: UNet, UNet++, DeepLab, and PSPNet, with ResNet50 and MobileNetV2 as encoders for polyp segmentation to decrease the chance of polyps being overlooked.

## 3 Methodology

### 3.1 Methods

The major steps in this project involved familiarizing ourselves with the Kvasir-SEG dataset, preprocessing the polyp and mask images, selecting and combining models, training, and evaluating the results.

### 3.2 Dataset

The Kvasir-SEG dataset (size 46.2 MB) contains 1000 polyp images and their corresponding ground truth from the Kvasir Dataset v2. The resolution of the images contained in Kvasir-SEG varies from  $332 \times 487$  to  $1920 \times 1072$  pixels. The images and its corresponding masks are stored in two separate folders with the same filename. The image files are encoded using JPEG compression, and online browsing is facilitated. The open-access dataset can be easily downloaded for research and educational purposes.

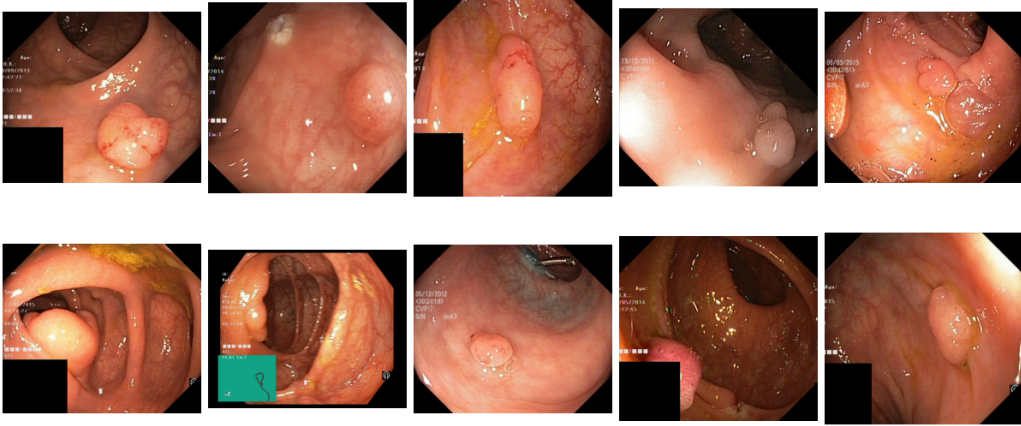


Fig2. Polyp samples

### 3.3 Encoders

The encoder subnetwork takes an input image and compresses it into a lower-dimensional representation, also known as a latent space. This process involves passing the input image through multiple layers of convolution and

pooling operations, which gradually reduce the spatial dimensions of the image while extracting important features Fig.2. In our models for implementing encoder-decoder networks, we should pick an appropriate encoder. We use some popular pre-trained encoders.

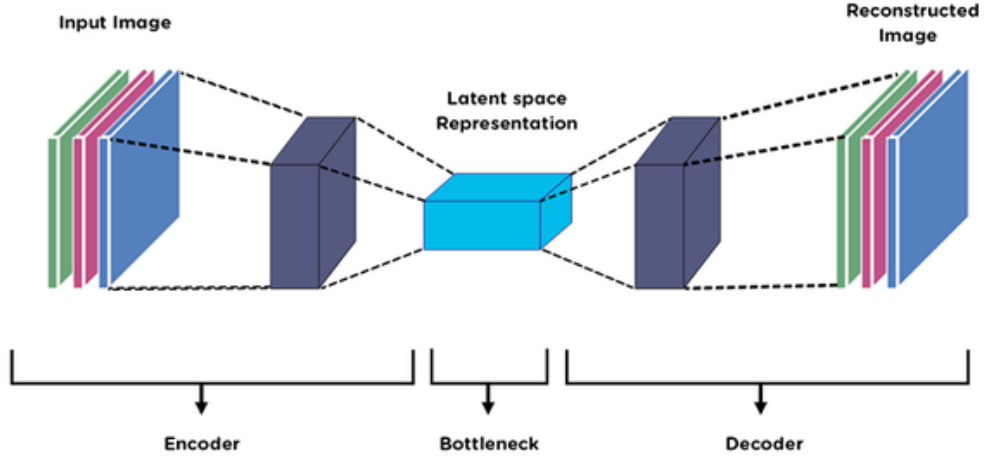


Fig.2 Encoder And Decoder Structure

### 3.3.1 MobileNet V2

MobileNet was created by Google and has a special layer called depthwise separable convolution, which is a block consisting of depthwise convolution and pointwise convolution. The purpose of this layer is to reduce computation (to have fewer parameters) so that a smaller model size can be obtained. MobileNet is a small, low-latency, low-power model measured to meet the resource constraints of various use cases. According to research papers, MobileNetv2 improves the performance of advanced mobile models on many tasks and benchmarks as well as across the entire spectrum of different model sizes. MobileNetv2 is a highly effective feature extractor for object detection and segmentation.

### 3.3.2 Resnet50

The ResNet50 is made up of 5 convolutional blocks on the 5 consecutive sizes of the encoding part. Each block is implemented via the Residual module that learns the difference between the input and the output of the block. This residue is obtained by passing the input value through 2 or 3 convolutional layers. the key innovation of ResNet (Residual Networks) lies in its use of residual connections, also known as skip connections. These skip connections allow the network to bypass one or more layers, enabling the flow of information directly from earlier layers to later layers. This addresses the vanishing gradient problem, which can occur in very deep neural networks during training. By using skip connections, ResNet enables the network to learn residual mappings instead of directly learning the desired underlying mapping. This approach facilitates the training of much deeper networks while still maintaining good performance. As a result, ResNet architectures have been highly successful in various computer vision tasks, such as image classification, object detection, and semantic segmentation.

## 3.4 Docoders

### 3.4.1 U-Net

before UNET we used a common fully convolutional network, Fig.3 shows this model structure.

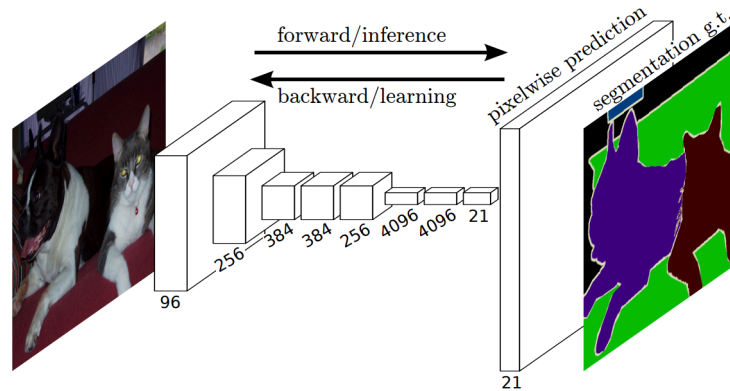


Fig.3 Fully convolutional networks can efficiently learn to make dense predictions for pixelpixel tasks like semantic segmentation.

UNET modify and extend this architecture such that it works with very few training images and yields more precise segmentations. The main idea in 3 is to supplement a usual contracting network by successive layers, where pooling operators are replaced by upsampling operators. Hence, these layers increase the resolution of the output. In order to localize, high resolution features from the contracting path are combined with the upsampled output. A successive convolution layer can then learn to assemble a more precise output based on this information. One important modification in UNET architecture is that in the upsampling part and UNET have also a large number of feature channels, which allow the network to propagate context information to higher resolution layers. As a consequence, the expansive path is more or less symmetric to the contracting path, and yields a u-shaped architecture. The network does not have any fully connected layers and only uses the valid part of each convolution, i.e., the segmentation map only contains the pixels, for which the full context is available in the input image. This strategy allows the seamless segmentation of arbitrarily large images by an overlap-tile strategy.

### 3.4.2 U-Net++

UNet++ introduces the concept of nested skip pathways to bridge this semantic gap. It adds additional skip connections between the encoder and decoder blocks at multiple resolutions. These connections allow the decoder to access and incorporate both low-level and high-level features from the encoder, providing a more detailed and comprehensive understanding of the image. UNet++ improved the traditional U-Net architecture by redesigning the skip connections and introducing a deeply supervised nested encoder-decoder network. The key difference between UNet and UNet++ lies in the architecture: UNet++ extends UNet by introducing nested skip pathways and dense skip connections, enabling more effective multi-scale feature aggregation and potentially improving segmentation accuracy.

### 3.4.3 PSPNet

The key idea and innovation behind PSPNet (Pyramid Scene Parsing Network) lie in the use of pyramid pooling modules to capture contextual information at multiple scales effectively. PSPNet's incorporates pyramid pooling

modules, which divide the feature maps into fixed-size grids and pool features from each grid separately. By doing so, PSPNet captures contextual information at multiple scales, enabling it to better understand the global context of the scene while preserving fine-grained details. This approach enhances the network’s ability to parse scenes accurately, especially in complex scenes with objects of various scales. Fig.4 Shows an overview of PSPNet.

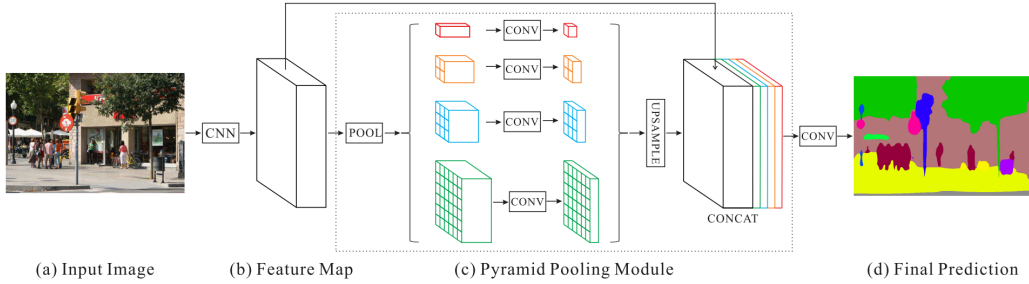


Fig.4 . Overview of PSPNet. Given an input image (a), we first use CNN to get the feature map of the last convolutional layer (b), then a pyramid parsing module is applied to harvest different sub-region representations, followed by upsampling and concatenation layers to form the final feature representation, which carries both local and global context information in (c). Finally, the representation is fed into a convolution layer to get the final per-pixel prediction (d).

#### 3.4.4 DeepLabV3Plus

DeepLabV3Plus achieves state-of-the-art semantic segmentation performance by leveraging atrous convolutions, encoder-decoder architecture with skip connections, and spatial pyramid pooling to effectively capture multi-scale contextual information while preserving spatial details.

### 3.5 Transfer learning

Deep convolutional neural network models may take days or even weeks to train on very large datasets. A way to shortcut this process is to reuse the model weights from pre-trained models developed for standard computer



vision benchmark datasets, such as the ImageNet image recognition tasks. Top-performing models can be downloaded and used directly or integrated into a new model for your own computer vision problems. In this project, I'll be using the UNet, UNet++, PSPNet, and DeepLabV3Plus architectures, with ResNet50 and MobileNet serving as encoders, to construct eight different encoder-decoder networks. These networks will leverage pre-trained weights from the ImageNet dataset.

## 4 Results

### 4.1 Metrics

#### 4.1.1 Dice Score

This competition is evaluated on the mean Dice coefficient. The Dice coefficient can be used to compare the pixel-wise agreement between a predicted segmentation and its corresponding ground truth. The formula is given by:

$$\frac{2 \times |X \cap Y|}{|X| + |Y|} \quad (1)$$

where X is the predicted set of pixels and Y is the ground truth. The Dice coefficient is defined to be 1 when both X and Y are empty. The leaderboard score is the mean of the Dice coefficients for each image in the test set.

In summary, the Dice score operates similarly to the F1-Score, measuring the agreement between predicted and ground truth regions. To normalize the score, we add 2 to the numerator. Figure 5 illustrates the Dice score in its simplest form.

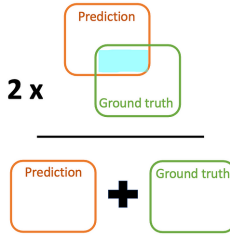
$$\text{Dice} = \frac{2 \times \text{Area of overlap}}{\text{Total area}} = \frac{2 \times \text{Prediction} \cap \text{Ground truth}}{\text{Prediction} + \text{Ground truth}}$$


Fig.5 Dice Score

#### 4.1.2 Dice Loss

Dice Loss is complement of Dice Score, So if you minimize the Dice loss you are maximizing Dice Score I used it as my loss function in training.

$$Dice - Loss = 1 - Dice - Score \quad (2)$$

#### 4.1.3 IOU Score

IOU is calculated by dividing the area of intersection between the predicted and ground truth regions by the area of their union. The formula for IOU can be expressed as follows:

$$\frac{2 \times |X \cap Y|}{|X \cup Y|} \quad (3)$$

A higher IOU value indicates a better alignment between the predicted and actual regions, reflecting a more accurate model. Figure 6 illustrates the Dice score in its simplest form.

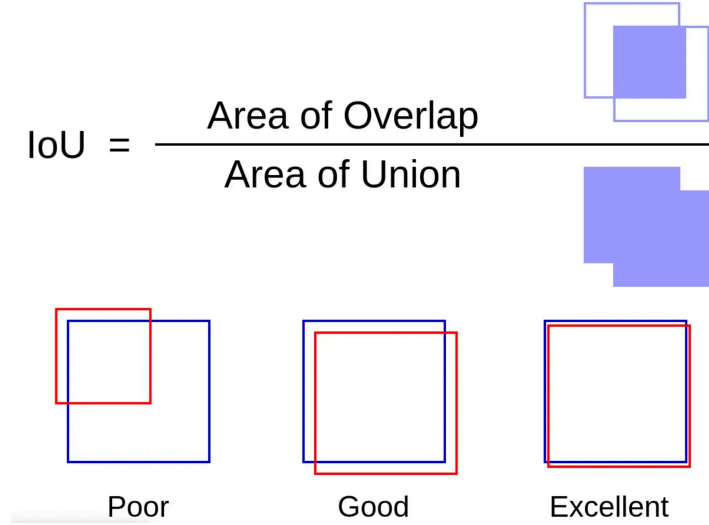
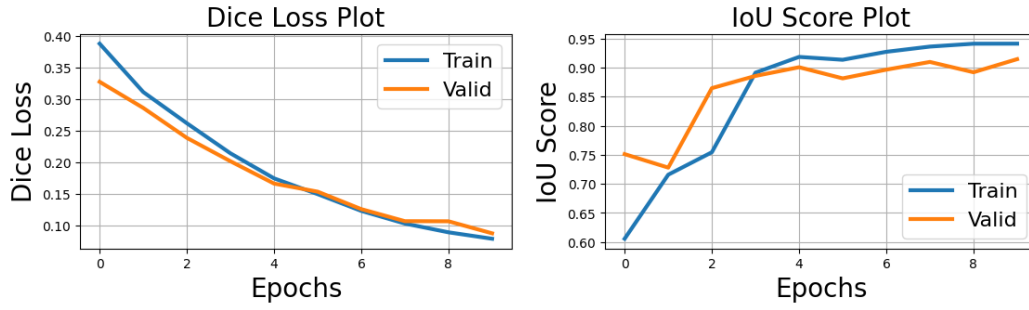


Fig.6 IOU Score

## 4.2 Training Results

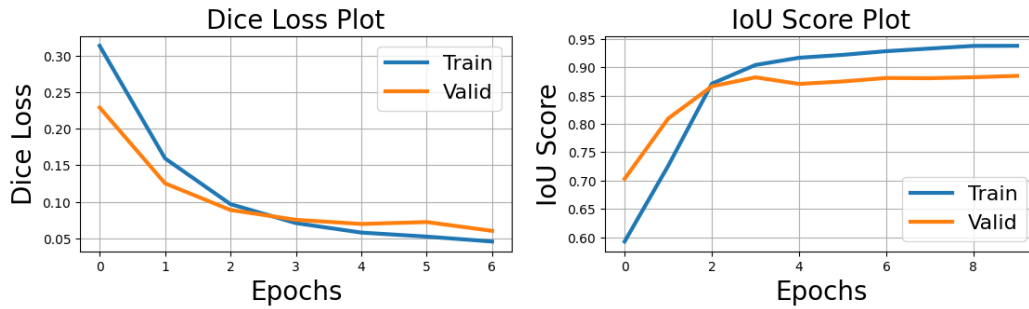
In this section, I will visualize the Dice Loss and IoU score for each model across training epochs to depict the training process of each model.



(a) Dice Loss

(b) IoU Score

Unet With Resnet50



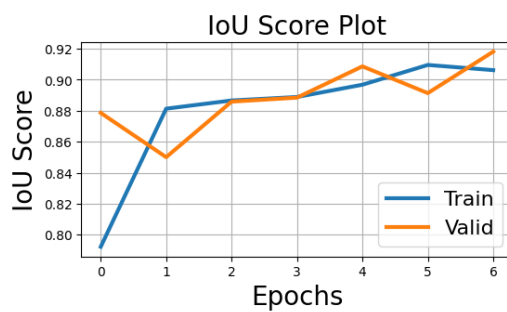
(a) Dice Loss

(b) IoU Score

Unet With MobileNetV2

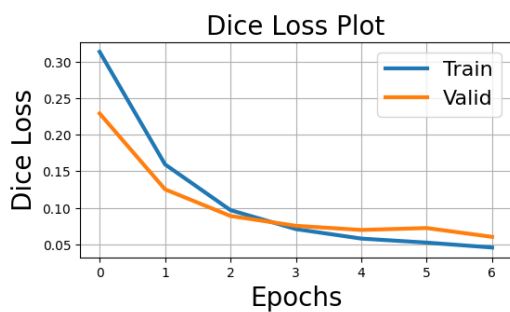


(a) Dice Loss

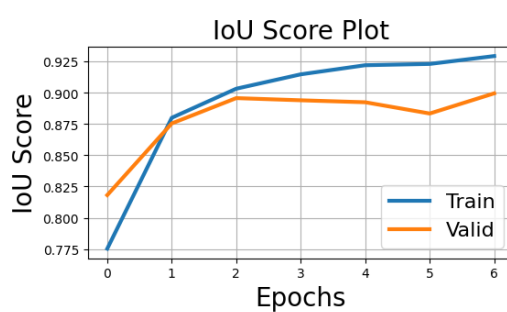


(b) IoU Score

Unet++ With Resnet50

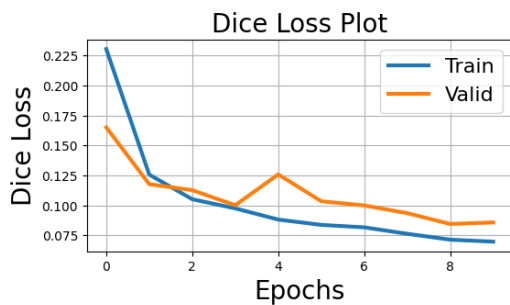


(a) Dice Loss

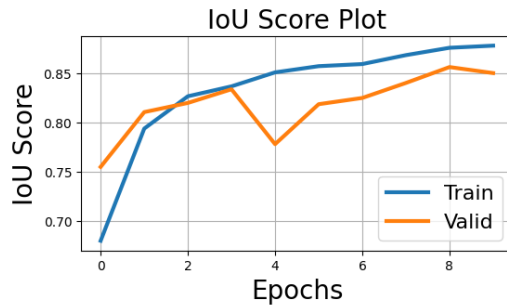


(b) IoU Score

Unet++ With MobileNetV2

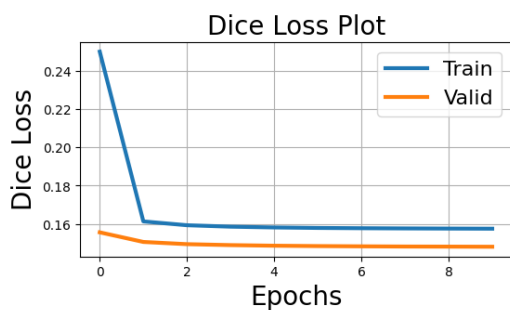


(a) Dice Loss

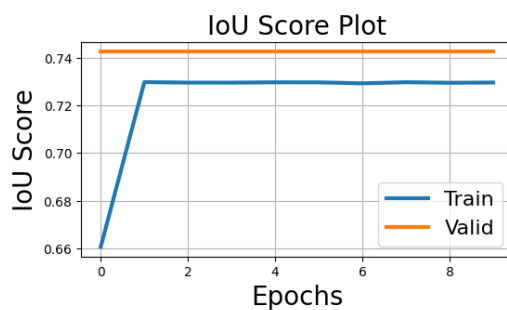


(b) IoU Score

PSPNet With Resnet50

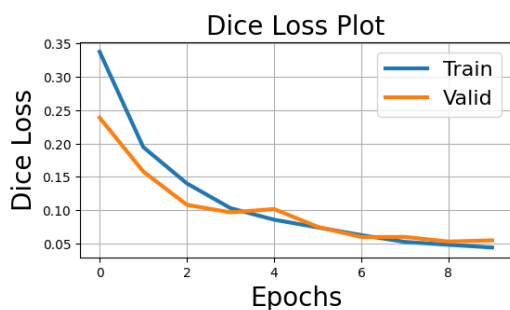


(a) Dice Loss

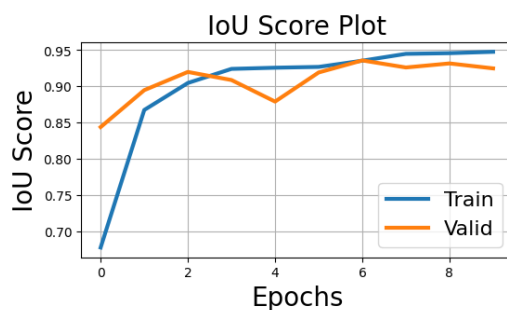


(b) IoU Score

### PSPNet With MobileNetV2

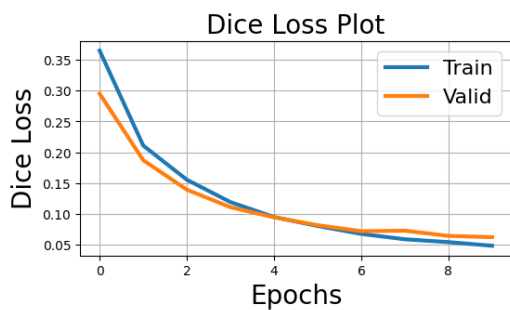


(a) Dice Loss

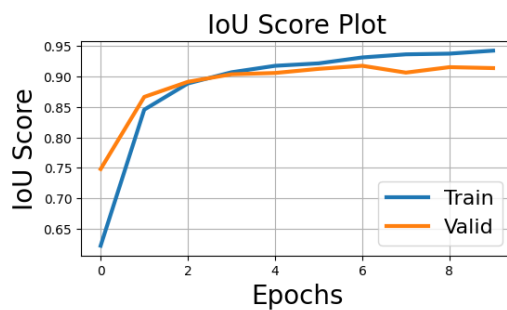


(b) IoU Score

### DeepLabV3Plus With Resnet50



(a) Dice Loss



(b) IoU Score

### DeepLabV3Plus With MobileNetV2

We understand that optimal training occurs when both the training and validation losses decrease across epochs while minimizing the gap between them. Additionally, the training and validation scores should increase with minimal disparity. In several models, such a pattern is observable. However, there are instances, like in the case of PSPNet with MobileNetV2, where training might not proceed optimally, potentially leading to suboptimal model performance.

### 4.3 Test Results

And Finally I run my test set on models and you can see the final result in Table 1.

	<i>ResNet50</i>			<i>MobileNetV2</i>		
	<i>IoU</i>	<i>DiceS</i>	<i>DiceL</i>	<i>IoU</i>	<i>DiceS</i>	<i>DiceL</i>
<i>UNet</i>	0.9261	0.9163	0.0837	0.9012	0.8972	0.1028
<i>Unet</i> + +	0.9229	0.9495	0.0505	0.9086	0.9418	0.0582
<i>PSPNet</i>	0.8652	0.9187	0.0813	0.7686	0.8586	0.1414
<i>DeepLV3</i>	0.9064	0.9316	0.0684	00.9264	0.9429	0.0571

**Table 1.** Performance analysis of different combinations of pretrained encoders and decoders on the Kvasir databaset.

Finally I plot the scores and losses for better visualization and reasoning.

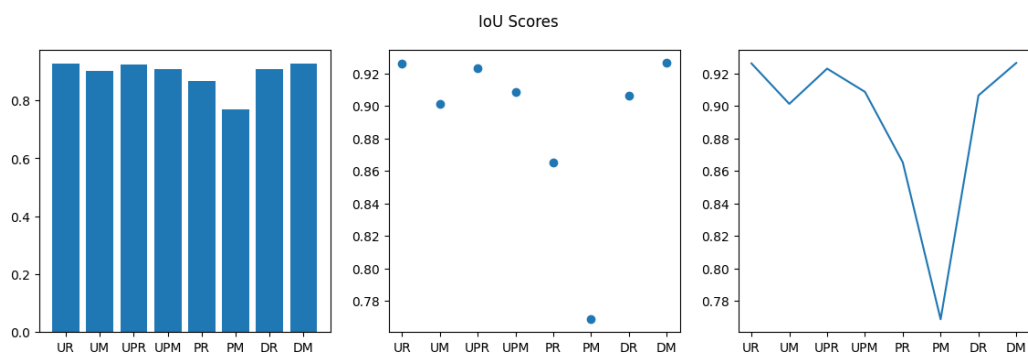


Fig.7 All Models IOU Scores

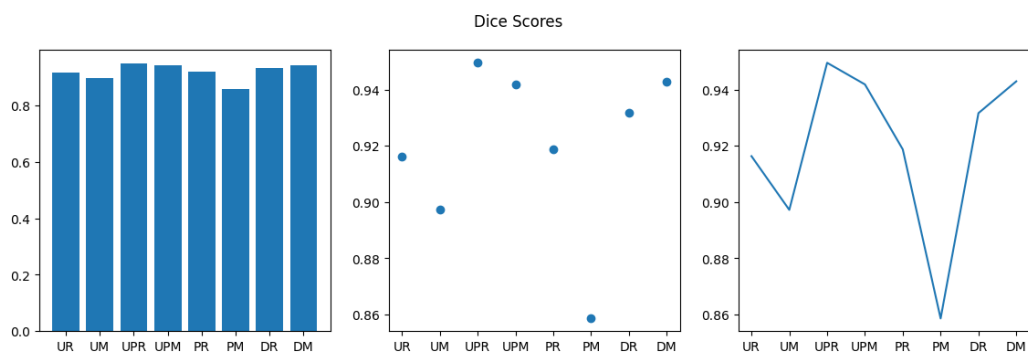


Fig.8 All Models Dice Scores

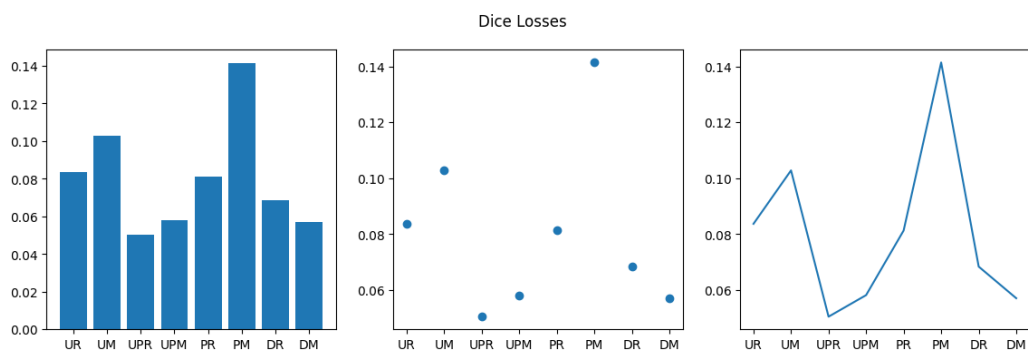


Fig.9 All Models Dice Losses

## 5 Conclusion

In this project we have analyzed multiple decoder-encoder networks for the task of polyp segmentation on endoscopic images. we combine these models to gain best result and finally state-of-the-art-model for this task is UNet++ with ResNet50.

## References

- [1] Double Encoder-Decoder Networks for Gastrointestinal Polyp Segmentation. Adrian Galdran , Gustavo Carneiro and Miguel A. Gonz´alez Ballester. 5 Oct 2021.
- [2] U-Net: Convolutional Networks for Biomedical Image Segmentation. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 18 May 2015.
- [3] Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Aug 22, 2018.
- [4] Pyramid Scene Parsing Network. Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, Jiaya Jia, The Chinese University of Hong Kong 2SenseTime Group Limited. 27 Apr 2017
- [5] Fully Convolutional Networks for Semantic Segmentation. Jonathan Long, Evan Shelhamer, Trevor Darrell. 8 Mar, 2015.
- [6] Segmentation of Polyps in Gastrointestinal (GI) Tract Images. Sabrina Nasrin, Javaneh Alavi, Pamila Viswanathan. 23 Dec 2021.
- [7] Image Segmentation of Normal Pap Smear Thinprep using U-Net with Mobilenetv2 Encoder. Deviana Sely Wita. 30June 23.
- [8] U-Net architecture variants for brain tumor segmentation of histogram corrected images. Sz. Lefkovits, L. Lefkovits. 2022.
- [9] Encoder–Decoder Convolutional Neural Networks for Flow Modeling in Unsaturated Porous Media: Forward and Inverse Approaches. Mohammad Reza Hajizadeh Javaran. 14 July 2023
- [10] CS231n: Convolutional Neural Networks for Visual Recognition