A Novel Approach to Medical Image Segmentation with Convformer-Based Attention Mechanism and UNet

A dissertation submitted in
partial fulfilment for the degree of

Master of Technology

in

Computer Science

by

Swastik Nandi
Roll no. - [**CS2234**]

under the supervision of

Dr. Swagatam Das
Professor
Electronics and Communication Sciences Unit (ECSU)

Indian Statistical Institute, Kolkata
June, 2024

CERTIFICATE

This is to certify that the dissertation titled, **"A Novel Approach to Medical Image Segmentation with Convformer-Based Attention Mechanism and UNet"**, submitted by **Mr. Swastik Nandi**, registration number CS2234, is a bonafide record of work carried out under my supervision. To the best of my knowledge, neither this dissertation nor any part of it has previously been submitted for the award of a degree, diploma or any other academic title anywhere else.

In my opinion, the dissertation fulfils all the requirements for the award of the degree of **Master of Technology in Computer Science.**

June,2024

Dr. Swagatam Das
Professor
Electronics and Communication Sciences Unit,
Indian Statistical Institute,
Kolkata-700108

Acknowledgement

I wish to express my gratitude to my supervisor, Dr. Swagatam Das for the opportunity provided to complete my dissertation. Foremost, I want to convey my utmost gratitude to Mrs. Susmita Ghosh, SRF, who helped me with every difficulty I faced and guided me with her knowledge of deep learning techniques and PyTorch coding.

I also acknowledge the computational resources provided to me which significantly facilitated my research work. I am thankful to my friends for their help, emotional support, and invaluable companionship. Additionally, I offer my deepest gratitude to my family for their support, encouragement, and comprehension during this endeavor..

Declaration of Originality

I, **Swastik Nandi**, with Roll No. **CS2234** hereby declare that the material presented in the dissertation titled

A Novel Approach to Medical Image Segmentation with Convformer-Based Attention Mechanism and UNet

represents original work carried out by me for the degree of **Master of Technology, Computer Science** at **Indian Statistical Institute, Kolkata.**
With my signature, I certify that:

- I have not manipulated any of the data or results.
- I have not committed any plagiarism of intellectual property. I have clearly acknowledged and referenced the contributions of others.
- I have understood that the work may be screened for any form of academic misconduct.

Swastik Nandi

Swastik Nandi
Roll no.- CS2234

Abstract

Accurate segmentation of medical images is a critical task in the field of healthcare, aiding in precise diagnosis and effective treatment planning. This project explores the enhancement of image segmentation models through the integration of advanced attention mechanisms. Our primary objective is to compare various attention techniques to develop a lightweight yet highly accurate model suitable for real-time applications. Given the significant body of work in medical image segmentation, our approach seeks to balance accuracy with computational efficiency. By incorporating different attention mechanisms and rigorously evaluating their performance, we aim to identify the optimal strategy for improving segmentation outcomes. The results demonstrate the potential for improved segmentation accuracy and efficiency, highlighting the effectiveness of attention-based models in capturing intricate patterns and dependencies within medical imaging data. We found out in our work that the CNN-based attention mechanism, or Convformer, effectively overcomes the issues related to the training conflict between CNNs and transformers. This project sets the groundwork for future advancements in semi-supervised and weakly-supervised learning, and we plan to expand our model's applicability across a broader range of medical imaging scenarios. Our ultimate objective is to contribute towards the development of robust, efficient, and adaptable segmentation models that can enhance diagnostic accuracy and patient care in various medical fields.

Keywords: Segmentation, Depth-wise Convolutions, Attention, Dice Score, Kvasir-Seg, ISIC2017, BraTS2020

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Image segmentation involves identifying a set of pixels with similar characteristics, where each pixel is assigned a label indicating its category. This process is essential in clinical usages such as diagnosis, treatment planning, and surgical interventions. It plays a vital role in delineating anatomical structures, tumors, lesions, or other regions of interest within medical images. Formally, medical image segmentation refers to the process of delineating boundaries of anatomical structures across various types of 2D/3D medical images. There are some critical challenges in this field of study. Firstly we have a data constraint issue, i.e. obtaining thousands of training images is typically unattainable. One of the most prominent reasons for it is the requirement of precise and accurate annotations which can only be performed by medical professionals such as radiologists, pathologists, or specialized technicians. Thus annotating these images accurately requires significant expertise, making the process labor-intensive and costly affair. Secondly, medical images contain sensitive patient information, and there are stringent laws and ethical guidelines around data sharing to protect and preserve patient privacy. This limits the availability of large, publicly available, and accessible datasets. Thirdly, images in the medical domain often have noisy/ill-defined boundaries, varying pixel intensities, and irregular shapes with significant variability(Figure 1.1 depicts the mentioned issues). Lastly, medical data is often isolated within individual hospitals, i.e. they purposely avoid sharing the data with other institutions making it difficult to aggregate large datasets from multiple sources.

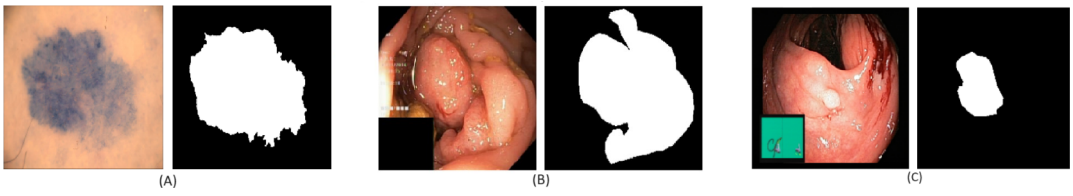

Figure 1.1: (A) depicts the noisy/ill defined boundary of skin lesion present in ISIC 2017[1] dataset. (B) and (C) depicts irregularity in shapes of tumours in Kvasir Seg polyp dataset[2]

Several traditional computer vision techniques such as thresholding [7], region growing [8], and active contours [9] have been historically employed for the task of image segmentation. However, these methods have some notable disadvantages. Thresholding methods are

highly sensitive to variations in lighting and contrast, making them unreliable for images with varying intensities. Similarly, the effectiveness of region-growing techniques heavily relies on the selection of initial seed points. Poorly chosen seeds can lead to sub-optimal segmentation results. Active contour models (snakes contour) require careful initialization and are sensitive to the initial contour placement and they often get stuck in local minima and fail to converge to the desired boundaries. However, the advent of deep learning has revolutionized this field, with deep learning-based models consistently outperforming classical techniques because of their capacity to autonomously extract high-level, abstract information from data through multiple hidden layers and their provision to get trained through multiple images, makes them robust to the variations in the input images. These models can identify intricate patterns and features from raw data which eliminates the need for manual feature extraction.

Convolutional Neural Networks (CNNs) have established themselves as the pioneering and most valuable models in the field of image segmentation because of their remarkable capacity to extract hierarchical features from raw pixel data. CNNs can autonomously identify complex patterns through multiple layers of convolution and pooling operations. This hierarchical feature extraction helps CNNs to capture low-level details like edges and textures in the initial layers, and simultaneously learn higher-level features like shapes in deeper layers. Convolutional layers, pooling layers, and fully connected layers in CNN architectures allow them to effectively handle the spatial hierarchies present in images. Convolutional layers apply a set of trainable kernels/filters to the input image, creating feature maps that highlight important aspects of the image. Pooling layers then downsample these feature maps, reducing their dimensions and retaining the most significant information, which helps in making the model invariant to small translations and distortions in the input data. Moreover, CNNs leverage large amounts of labeled data and powerful computational resources to optimize their performance through backpropagation and gradient descent techniques. This training process enables CNNs to fine-tune their filters' weights, resulting in highly accurate and robust segmentation models. Additionally, advancements in network architectures, such as U-Net[10], Mask RCNN[11], and Fully Convolutional Networks (FCNs)[12], have specifically tailored CNNs for image segmentation tasks, further enhancing their effectiveness.

Among all the CNN models used for image segmentation, U-Net[10] is the most popular model in the realm of medical image segmentation due to its architectural brilliance. U-Net features a symmetric encoder-decoder structure that effectively captures contextual details at multiple scales. Along with this U-Net possesses the most advantageous feature which is the concept of skip connections. Connecting corresponding layers in the encoder and decoder paths, skip connections concatenate feature maps from the encoder to the decoder, allowing the network to preserve high-resolution features essential for precise segmentation. This helps U-Net to combine the low-level data (details) with high-level data (context), leading to more accurate and detailed segmentation. Along with this, attention mechanism[5] also comes into the picture in the domain of medical image segmentation due to its ability to capture long-range dependencies and concentrate on critical areas of an

image, enhancing the detection and segmentation of small or subtle regions such as lesions or tumors. This capability improves boundary accuracy, which is essential for precise medical diagnosis by accurately delineating complex structures with very fine boundaries. Attention mechanisms are also accustomed to managing multi-scale features effectively, enhancing segmentation capabilities across different object sizes within medical images.

Deep learning models offer numerous advantages, but they come with a few notable drawbacks. One significant challenge is the slow training process, attributed to the extensive parameter count. This necessitates high-performance computational resources like expensive GPUs to expedite training. This in turn presents before us a challenge to balance between cost and performance. To mitigate this issue, there is a need for lightweight models that uphold high performance standards keeping the parameter count as less as possible. Additionally, the substantial parameter count with limited training data can lead to the problem of over-fitting, thus resulting in poorer performance on test data.

In this study, our focus lies in exploring different attention techniques, presenting a comparative study of their efficiency, and proposing a lightweight model that has good efficacy in terms of both parameter count as well as accuracy. Leveraging U-Net as our backbone network, we assess our models across a range of diverse datasets, spanning polyp, skin lesion, and brain tumor datasets. These datasets are carefully chosen to test the robustness of our models. Our primary aim is to maintain a harmonious balance between model intricacy and performance, striving for efficient training without compromising accuracy significantly.

1.1 Formulation of Image Segmentation Problem

Let us take an image $X \in \mathbb{R}^{H \times W \times C}$ where H, W, and C are the height, width and number of channels (e.g., 3 for RGB images) respectively. The objective of image segmentation is to assign a label l_i from a set of possible labels $\{1, 2, \ldots, L\}$ to each pixel i in the image. This can be expressed mathematically as:

Define a segmentation function $f : \mathbb{R}^{H \times W \times C} \rightarrow \mathbb{R}^{H \times W \times L}$ such that:

$$\hat{Y} = f(X)$$

where $\hat{Y} \in \mathbb{R}^{H \times W \times L}$ is the output segmentation map, and each pixel i in $\hat{Y}$ contains a vector of length L representing the probability distribution over the L possible labels. For a task of binary segmentation, the number of output labels is two which indicates the foreground and background pixels, and a multi-class segmentation task outputs more than two labels which indicates the different clusters present in the output map.

We have the ground truth segmentation map denoted by $Y \in \mathbb{R}^{H \times W \times L}$ and the predicted segmentation map denoted by $\hat{Y}$.

The loss, denoted as $\mathcal{L}$, is computed as the mismatch between the predicted segmentation map $\hat{Y}$ and the actual segmentation map Y.

$$L = \text{Loss}(Y, \hat{Y})$$

where Loss can be Binary Cross Entropy(BCE) Loss, Mean Squared Error(MSE) Loss, Dice Loss etc.

During the training process, the model parameters are adjusted iteratively to minimize this loss. The aim is to improve the accuracy of the segmentation model by making the predicted segmentation map $\hat{Y}$ as identical as possible to the ground truth segmentation map Y. Now we provide a general algorithm for the training process involved in medical image segmentation task which we have used to train our models as discussed in the later sections.

Algorithm 1 Training Model for Medical Image Segmentation

```
Initialize the model architecture, loss function, evaluation metric, learning rate and optimization algorithm.
Split dataset into training and test sets.
Pre-process images and corresponding segmentation maps with suitable transformations.
repeat
    for every batch in training set do
        Forward pass: Calculate predicted segmentation masks.
        Calculate loss between predicted and ground truth masks.
        Backward pass: Calculate gradients and update model parameters accordingly.
    end for
    Evaluate model on test set and save model based on highest obtained test score.
until Number of epochs is completed
Evaluate final model on test set and report the score based on the evaluation metric.
```

1.2 Objective

The objective of this thesis work is two-fold: first, explore various attention-based convolutional neural network (CNN) models(more specifically attention-based U-Net models) for medical image segmentation, and secondly, to develop lightweight architectures to address computational complexities and challenges in deep learning-based segmentation. The exploration of attention mechanisms involves analyzing the effectiveness of self-attention, spatial attention, channel attention, and convolution-based self-attention models in enhancing CNNs' ability to capture relevant spatial and contextual information within medical images. By integrating these mechanisms into U-Net architecture, the aim is to improve the accuracy and robustness of segmentation across various medical imaging datasets. Simultaneously, the development of lightweight models focuses on optimizing model complexity and parameter efficiency to minimize computational resource requirements. The objective aims to facilitate efficient training of segmentation models in resource-constrained clinical environments. Together, these objectives seek to present a comparative study on various kinds of attention techniques and simultaneously propose an alternative lightweight architecture that is efficient both accuracy-wise and parameters-wise.

1.3 Related Work

Numerous related works in medical segmentation have employed U-Net as the foundational architecture. One notable example is UNet++[13] which is a deeply supervised encoder-decoder network that features dense, nested skip pathways between the encoder and decoder sub-networks. These redesigned skip connections intend to minimize the semantic disparity between the encoder and decoder feature maps. To address issues in UNet++, UNet 3+[14] was introduced, incorporating full-scale skip connections and deep supervision. These full-scale skip connections blend high-level semantics with low-level details from feature maps across different scales, while deep supervision facilitates learning hierarchical representations derived from the combined feature maps.

Another advancement is UNeXt[15], "A Convolutional multilayer perceptron (MLP) based network for image segmentation". UNeXt features an initial convolutional stage followed by an MLP stage in the latent phase, with an efficiently tokenized MLP block.

In addition to these, V-Net[16] was introduced for 3D image segmentation, utilizing a volumetric, fully convolutional neural network. To mitigate the high computational demands of 3D convolutions, H-DenseUNet[17] was proposed. This model integrates a 2D DenseUNet for efficient feature extraction within individual slices and its three-dimensional equivalent to hierarchically aggregate volumetric contexts, particularly for liver and tumor segmentation.

ResUNet-a[18], a residual learning U-Net model, combines a U-Net architecture incorporating an encoder/decoder backbone, along with atrous convolutions, residual connections, pyramid scene parsing pooling, and multitasking inference capabilities.

Recently, transformer-based networks have gained prominence in medical image segmentation due to their ability to capture global image contexts. ViT[19], a model initially designed for image classification, uses a Transformer-like architecture over image patches. This approach divides an image into patches of fixed size, followed by linearly embedding them with position embeddings, and feeding the resulting sequence into a conventional Transformer encoder. ViT marked one of the earliest applications of transformers in medical image segmentation.

Further developments include TransUNet[20], which combines a hybrid CNN-Transformer architecture designed to harness detailed high-resolution spatial information from CNN features alongside the global context encoded by transformers. MedT[21] addresses data constraints in medical datasets by introducing a gated axial-attention model, enhancing self-attention mechanisms with an additional control layer. MedT also introduces a Local-Global training strategy (LoGo) aimed at enhancing performance.

Prior Attention Network (PANet[22]) employs a a coarse-to-fine strategy for segmenting multiple lesions in medical images. PANet integrates a lesion-related spatial attention mechanism within the network and utilizes intermediate supervision to generate lesion-related attention, thereby accelerating convergence and enhancing segmentation performance.

These advancements underscore the ongoing evolution of medical image segmentation models through the integration of sophisticated attention mechanisms and transformer

architectures with traditional convolutional neural networks (CNNs). However, these advanced techniques often result in a significant increase in the parameter count. For example, the base ViT model[19], a benchmark in this field, contains 86 million parameters, which is notably high. Therefore, our objective was to reduce the parameter count without substantially compromising the accuracy of the model.

The outline of the dissertation is as follows: chapter 2 explains in detail the working of U-Net, depth-wise convolutions and DW U-Net. It also presents why DW U-Net can be used as a substitute of U-Net with experiment details and results. chapter 3 presents different attention techniques and their results on three datasets and finally chapter 4 presents the concluding remarks and future direction of work.

Chapter 2

U-Net and Depth-wise Convolution

This chapter primarily delves into the core structure of our project, focusing on two key components: the U-Net architecture and depth-wise separable convolutions. The following sections will provide an in-depth understanding of these topics.

2.1 U-Net

U-Net is a commonly employed and tested deep learning model that was initially introduced in the paper "U-Net: Convolutional Networks for Biomedical Image Segmentation" [10]. Addressing the challenge of limited annotated data in the medical area of research was the main purpose of this architecture. This network was designed to address these challenges by efficiently utilizing a smaller dataset while preserving both speed and accuracy.

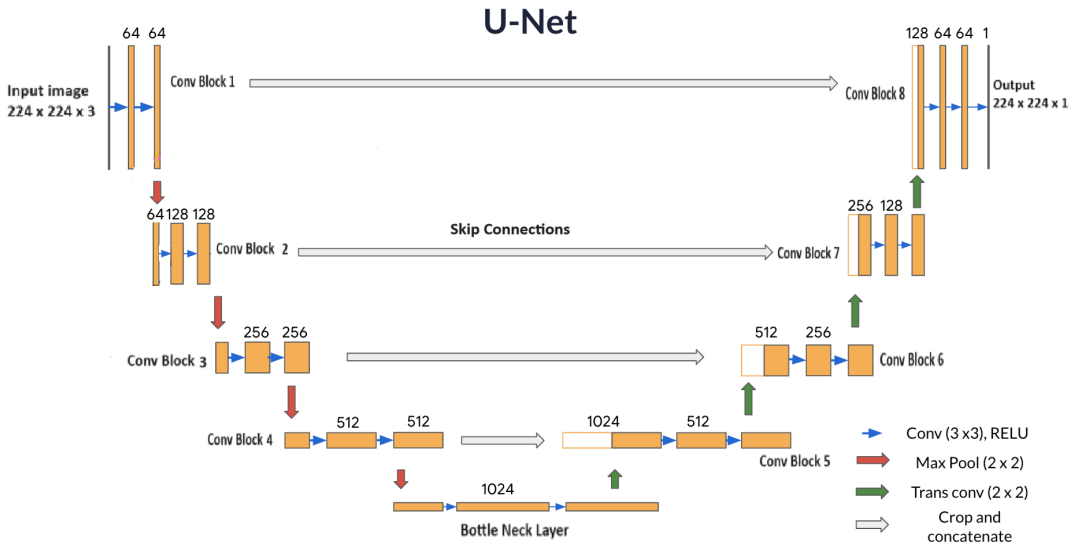

Figure 2.1: U-Net architecture

2.1.1 Architecture of U-Net

Overview

The framework of U-Net [10] is uniquely designed with both an expansive and a contracting path. The encoder layer forms the contracting path, which reduces the spatial dimension of the input while capturing contextual information. The expansive path, containing a decoder layer, decodes the encoded data, upsamples the data and utilizes the information extracted from the contracting path with the help of skip connections to generate a segmentation map. The contracting path in U-Net has the responsibility to identify pertinent features within the input image. The encoder layers perform convolutional and pooling operations that reduce the spatial resolution of the feature maps while increasing their depth, i.e. they decrease the spatial dimensions and increase the number of channels, thus capturing progressively abstract representations of the input. This contracting path is similar to the feed-forward layers in other convolutional neural networks. Conversely, the expansive path has the responsibility to decode the encoded data and locate its' features while maintaining the spatial resolution of the input. The decoder layers up-sample the feature maps, performing transpose convolutional operations. The skip connections from the contracting path to the expansive path aid in retaining the spatial information that is lost in the contracting path, enabling the decoder layers for more precise feature localization.

Detailed Explanation

Figure 2.1 explains in detail the overall architecture of U-Net. In the figure, convolution blocks Conv Block 1, Conv Block 2, Conv Block 3 and Conv Block 4 represent convolution blocks in the encoder layer and Conv Block 5, Conv Block 6, Conv Block 7 and Conv Block 8 represent convolution blocks in the decoder layer. We give input of dimension $224 \times 224 \times 3$ to Conv Block 1 and we get output of $224 \times 224 \times 1$ as output from Conv Block 8. In each convolution block of the encoder layer, we do two successive convolution operations with kernel size 3×3, padding 1 and stride 1. At the end of two convolutions in one block, the pooling operation (Max Pool with kernel size 2×2) is done and the feature map is passed onto the next Conv Block. For example, in the first convolution block $224 \times 224 \times 3$ is the input after two convolutions the feature map becomes $224 \times 224 \times 64$, and then pooling is done to make the dimension $112 \times 112 \times 64$ (as the kernel size is 2×2 thus the dimension gets divided by 2). Subsequently, the other 3 convolution blocks are executed and the dimension of the feature map becomes $14 \times 14 \times 512$ before entering the bottleneck layer. In the bottleneck layer also there are 2 convolution operations and at the end of it, we get the feature map of dimension $14 \times 14 \times 1024$. After that transpose convolution operation is performed which makes the dimension of the feature map $28 \times 28 \times 512$. With this feature map, the feature map from Conv Block 4 is concatenated with the help of a skip connection. In each block of the decoder section after up-convolution and concatenation through skip connections, two convolution operations are successively performed with kernel size 3×3, padding 1 and stride 1. Thus the output at the end

of Conv block 5 has dimension $28 \times 28 \times 512$. Subsequently, three other blocks of the decoder layer are executed and finally, we have an output of dimension $224 \times 224 \times 1$.

2.2 Depth-wise Separable Convolutions

Convolutions are of many types. One of the most important types of convolution is the depth-wise separable convolution. The major advantage of depth-wise separable convolutions is that they have fewer parameters to adjust as compared to normal convolutions which in turn reduces the chances of overfitting. Also, they are computationally cheaper due to the lesser number of computations involved in the process.

2.2.1 Working of Depth-wise Separable Convolutions

Convolution Operation

Let us take an input of dimension $D_f \times D_f \times M$, where $D_f \times D_f$ represents the image spatial dimensions and M denotes the number of channels present(3 for an RGB image). Suppose there are N kernels whose size is $D_k \times D_k \times M$. According to convolution mathematics, if a convolution operation is performed, the output size will be $D_p \times D_p \times N$.

The number of multiplications in one convolution operation is equal to $D_k \times D_k \times M$, which is actually the size of the filter.

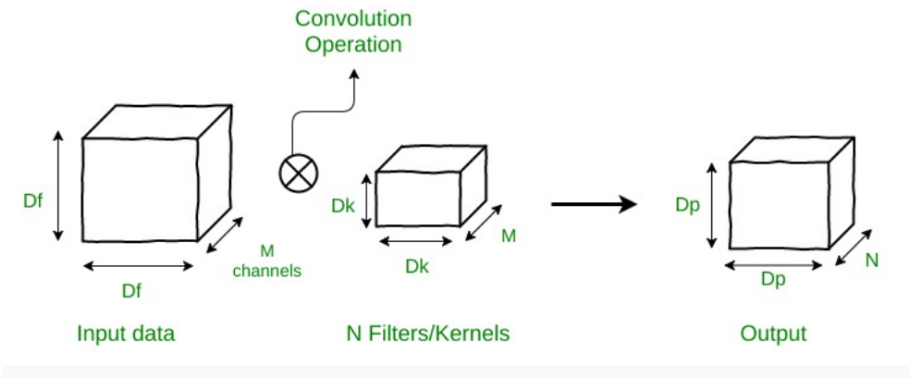

Figure 2.2: Working of Convolution operation[3]

Now, we have N filters, and each of them slides in both the vertical and horizontal directions D_p times. Thus the total number of multiplications in such case becomes $N \times D_p \times D_p \times$ (multiplications per convolution). Therefore, total number of multiplications for a standard convolution operation is given by:[3]

$$\text{Total number of multiplications} = N \times D_p^2 \times D_k^2 \times M.$$

Depth-Wise Separable Convolutions

When considering depth-wise separable convolutions, the process can be divided into two distinct steps:

1. Depth-wise convolutions
2. Point-wise convolutions

- In the depth-wise convolution step, the convolution operation is applied to each channel separately, in contrast to the standard CNNs which apply it across all M channels simultaneously. Thus, the convolutional kernels will be two dimensional and have dimensions $D_k \times D_k \times 1$. Given the input data has M channels, we will need M such 2D filters. Consequently, the output will have dimensions $D_p \times D_p \times M$. Each individual convolution operation involves $D_k \times D_k$ multiplications. As the filter slides over the input $D_p \times D_p$ times for all M channels,
  Thus total number of multiplications for the depth-wise convolution $= M \times D_k^2 \times D_p^2$.

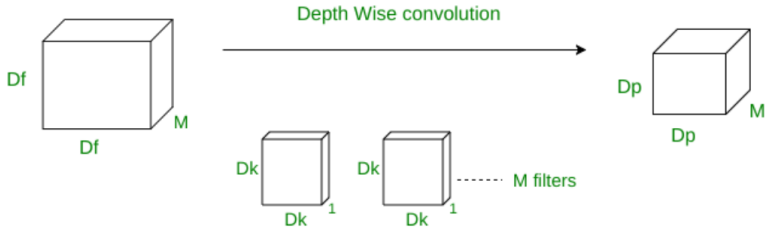

Figure 2.3: Mechanism of Depth-Wise Convolution[3]

- In point-wise convolution step, a 1×1 convolution is applied across all the M channels. The filters' dimension for this operation will be $1 \times 1 \times M$. If we use N such filters, the output will have dimensions $D_p \times D_p \times N$. Each point-wise convolution operation thus requires $1 \times M$ multiplications. As the filter is applied $D_p \times D_p$ times, Then total number of multiplications for the point-wise convolution $= D_p^2 \times M \times N$.

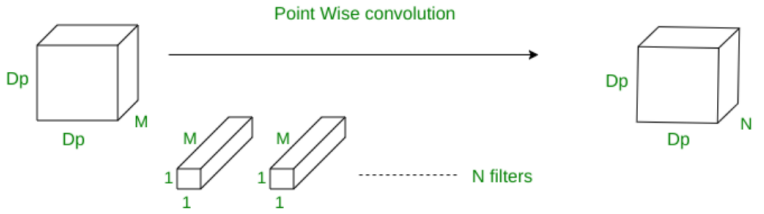

Figure 2.4: Mechanism of Point-Wise Convolutions[3]

Therefore, for the complete depth-wise separable convolution operation,

Total multiplications = Depth-wise convolution multiplications + Point-wise convolution multiplications

$$\begin{aligned}\text{Total multiplications} &= M \times D_k^2 \times D_p^2 + D_p^2 \times M \times N \\ &= D_p^2 \times M \times (D_k^2 + N)\end{aligned}$$

$$\frac{\text{Computational cost of depth-wise separable convolutions}}{\text{Computational cost of standard convolutions}} = \text{RATIO (R)}$$

For instance, considering $N = 100$ and $D_k = 512$, we find the ratio R to be 0.01.

This implies, in this example, the depth-wise separable convolution method performs 100 times fewer multiplications compared to a conventional convolutional method.

2.3 DW U-Net

Now we propose a novel form of the U-Net architecture that is both lightweight and maintains high accuracy. Traditional U-Net models employ standard 2D convolutions, which, although effective, result in a large number of parameters, leading to increased chances of overfitting, large computational costs, and higher memory usage. Our approach seeks to mitigate these drawbacks by incorporating depth-wise separable convolutions, which consist of depth-wise convolution operation followed by point-wise convolution operation, in place of the standard 2D convolution operations. Thus, in our proposed architecture the basic framework of the U-Net model remains the same. Only in place of two standard 2D convolutions in each convolution block of both encoder and decoder layers(Figure 2.1), two depth-wise separable convolutions are incorporated. Apart from this, other architectural designs of this model are similar to the original U-Net model.

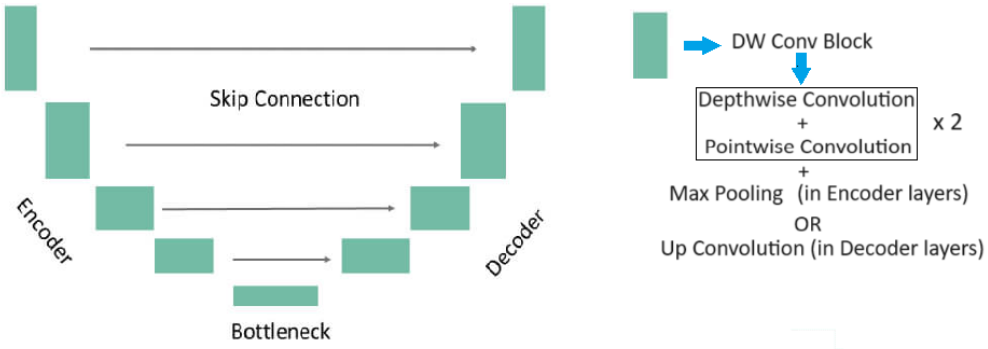

Figure 2.5: Block diagram of DW U-Net. Here each DW Conv Block contains two depth-wise convolution operation where each of them is succeeded by point-wise convolution. Apart from that the basic working of the architecture remains same as explained in subsection 2.1.1.

Table 2.1: Parameter comparison between U-Net and DW U-Net. Note that parameter count is based on 3 encoding/decoding and bottleneck layers.

Models	Number of Parameters
U-Net	31.03 M
DW U-Net	1.49 M

From Table 2.1 it is clear that this modification significantly decreases the number of parameters. Now we claim that our proposed model DW U-Net achieves comparable accuracy to conventional U-Net with lesser number of parameters. To establish this claim, we

conducted experiments whose detailed explanation is provided in the subsequent sections.

2.4 Experiment and Analysis

To validate our claim that the proposed DW U-Net (Depth-Wise U-Net) achieves comparable accuracy to the conventional U-Net with fewer parameters, we carried a series of experiments. The detailed methodology and results of these experiments are presented in the following subsections. The first experiment rigorously compares the performance and efficiency of DW U-Net and the original U-Net, substantiating our claim that DW U-Net is an effective and efficient alternative for medical image segmentation tasks. The second experiment involves an empirical study to determine the optimal number of layers needed in the DW U-Net structure.

2.4.1 Dataset

For this experiment, we utilized the Kvasir-SEG polyp dataset[2] (Segmented Polyp Dataset for Computer-Aided Gastrointestinal Disease Detection). Pixel-level image segmentation is of high demand in the domain of medical image analysis. It is challenging to obtain annotated medical images with corresponding segmentation masks due to the requirement for domain-specific expertise, which is typically possessed only by highly trained doctors and medical practitioners. Kvasir-SEG is an free-access dataset containing images of gastrointestinal polyp and their corresponding segmentation masks. To ensure the standard and authenticity of the dataset, these masks are manually annotated and verified by an experienced gastroenterologist.

The human gastrointestinal (GI) tract comprises various sections, one of which is the large bowel. This section can be affected by several abnormalities and diseases, such as colorectal cancer. Colorectal cancer is the second most common cancer type where polyps are the precursors. They are found in nearly half of individuals at the age of 50 who undergo screening colonoscopy, with incidence increasing with age. Colonoscopy is the gold standard for detecting and assessing these polyps, followed by biopsy and removal. Early detection of disease significantly impacts survival rates from colorectal cancer, making polyp detection crucial. Several studies have shown that there is a tendency of polyps being often overlooked during colonoscopies, with miss rates ranging from 14% to 30% depending on the type and size of the polyps. Increasing the accuracy of polyp detection significantly reduces the risk of colorectal cancer. Therefore, automatic detection of polyps at an early stage is necessary for both the prevention and survival rates of colorectal cancer. This serves as the main motivation and utility of the Kvasir-SEG dataset.

The Kvasir-SEG dataset comprises 1000 images of polyps and their corresponding ground truth masks, with resolutions ranging between 332×487 and 1920×1072 pixels. Thus, we selected the Kvasir-SEG dataset for our experiment to measure the robustness of our proposed DW U-Net model and provide an unbiased comparison study of U-Net and DW U-Net.

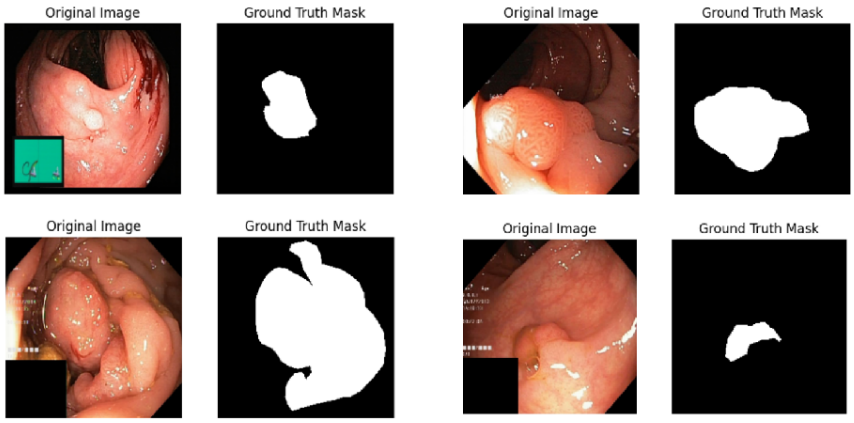

Figure 2.6: Examples of images and corresponding masks from Kvasir-SEG polyp dataset[2]

2.4.2 Experiment Details

For our experiment, we used the Kvasir-SEG[2] polyp dataset. This dataset comprises of 1000 images which is divided into train and test datasets. Upon splitting, we resized both the image and its corresponding mask to dimensions 224×224 in both the training and test datasets. As loss function we used Binary Cross Entropy(BCE loss) and Dice Score as the evaluation metric and ADAM[26] as the optimizer. The training process was executed for 100 epochs and the dice score was calculated on the test set.

Loss Function

We used Binary Cross Entropy(BCE) Loss as the loss function. The BCE Loss is a combination of a Sigmoid layer and the Binary Cross-Entropy Loss in one single class. This loss is more stable than a plain Sigmoid followed by a Binary Cross-Entropy Loss. The formula for the BCE Loss is given by:

$$\mathcal{L}_{\text{BCE}}(x, y) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log\left(\sigma(x_i)\right) + (1 - y_i)\log\left(1 - \sigma(x_i)\right)\right],$$

where x_i are logits, y_i are target labels, N is the number of samples and

$$\sigma(x_i) = \frac{1}{1 + \exp(-x_i)}$$

Evaluation Metric

We utilized the Dice Score as our evaluation metric, a widely recognized measure in computer vision and medical imaging for assessing the overlap or similarity between two sets. This metric is particularly common in image segmentation tasks, where it evaluates how accurately the predicted region matches the ground truth region, reflecting the actual location of objects within the image.

The Dice Score which is also known as the Sørensen–Dice coefficient is evaluated using

the following formula.

$$\text{Dice Score} = \frac{2 \times \text{Area of Intersection}}{\text{Total Area of Both Sets}}$$

Figure 2.7 gives us a visual representation of Dice score. Let C1 in green be the predicted region and C2 in red be the ground truth region. Then area of intersection refers to the overlapping elements(pixels) i.e. the region that belongs to both C1 and C2 and the total area of both sets refers to the sum of the number of pixels belonging to sets C1 and C2. The value of Dice Score lies between 0 and 1 where 0 indicates that there is no overlap between the two sets, meaning the predicted mask has no common elements with the ground truth and a score of 1 indicates perfect overlap, where the predicted mask is identical to the ground truth mask. A higher Dice Score indicates better segmentation accuracy. In medical imaging tasks, precise object delineation is critical. Thus we selected Dice Score as our evaluation metric as it can assess how well our model can capture the spatial information and boundaries of objects within the images.

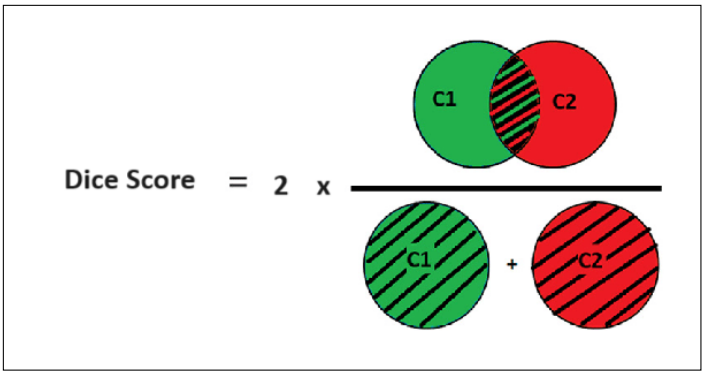

Figure 2.7: Pictorial representation of Dice Score

2.4.3 Results and Discussion

We conducted two experiments and both of them were conducted on the Kvasir-SEG[2] polyp dataset. Firstly to study the comparison of U-Net and DW U-Net and assess whether DW U-Net can be an efficient alternative to the standard U-Net architecture and secondly to present a comparison study for determining the optimal number of layers for DW U-Net. Please note that in the first experiment, we kept three encoding and decoding layers with one bottleneck layer in the architecture of both U-Net and DW U-Net.

For both experiments, we made sure that the data was well-shuffled before dividing it into train and test datasets. Shuffling prevents biases like patient-level and temporal bias and enhances the model's generalization ability. For both experiments, we did not use any transformation other than resizing both the image and its' corresponding mask to dimensions 224×224 and also ensured that the training conditions were similar for all the models. This was purposefully done such that no external factors could affect the results of the experiments. Both experiments were done to find the most appropriate backbone for our upcoming models which are defined in the subsequent chapters. Below are the results

of these experiments. Note that, these results are based considering 3 encoding/decoding and 1 bottleneck layer in both the cases of U-Net and DW U-Net.

Table 2.2: Comparison of U-Net and DW U-Net

Models	Number of parameters	Execution time	Dice Score
U-Net	31.03 M	62 mins	0.6502
DW U-Net	1.49 M	45 mins	0.6413

From Table 2.2, we have made several key observations. Firstly, the DW U-Net has significantly fewer parameters, approximately 20 times less than the original U-Net. This substantial reduction in parameters translates to a more lightweight model, which is advantageous for computational efficiency. In addition to it, the execution time of DW U-Net is about 1.4 times faster than that of the U-Net which further highlights its efficiency in terms of speed.

Despite these reductions, the Dice Score of DW U-Net is only slightly lower than that of U-Net, with U-Net achieving a score just 1.01 times higher. This minimal difference in performance indicates that the DW U-Net maintains a high level of accuracy in this particular segmentation task, even with its reduced complexity. These results indicate that the reduction in the number of parameters does not significantly deteriorate the Dice Score. Therefore, we can confidently conclude that DW U-Net is an effective alternative to the original U-Net, offering similar segmentation performance with the added advantages of lower computational requirements and faster execution times.

Table 2.3: Comparison study to determine the optimal number of layers for DW U-Net. Kindly note that the channel progression through layers is also mentioned in brackets.

Number of layers	Number of parameters	Dice Score
2 encoding/decoding and 1 bottleneck (64 - 128 - 256)	0.37 M	0.5544
3 encoding/decoding and 1 bottleneck (64 – 128 – 256 – 512)	1.49 M	0.6413
3 encoding/decoding and 1 bottleneck (32 - 64 – 128 – 256)	0.38 M	0.6146
4 encoding/decoding and 1 bottleneck (64 – 128 – 256 – 512 - 1024)	5.99 M	0.6571
5 encoding/decoding and 1 bottleneck (64 – 128 – 256 – 512 - 1024 - 2048)	23.88 M	0.6853

Table 2.3 reveals an important trend: upon increasing the number of layers in the model, the Dice Score tends to get enhanced thereby improving the segmentation accuracy. This improvement, however, results in a corresponding increase in the number of parameters. A higher number of parameters can lead to increased computational demands and potentially long training duration, which might not be an ideal situation for all applications. Also, our main objective is to find out an appropriate backbone for our subsequent models. Thus choosing a model with more parameters will eventually increase the chances of overfitting. Please note that the figures shown in brackets for each layer mentioned in Table 2.3 represent the progression of the number of channels in the encoder block of the

DW U-Net.

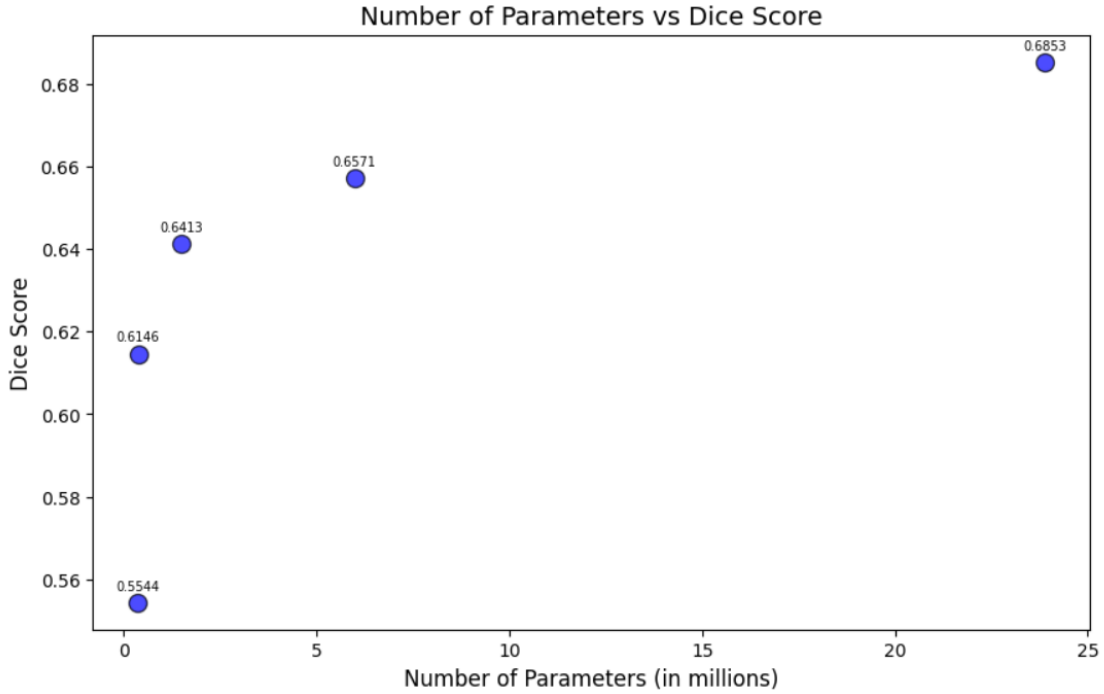

Figure 2.8: Scatter Plot of Number of Parameters vs Dice Score

Form Figure 2.8 we see that the model having Dice Score 0.6571 balances the trade-off between performance and number of parameters. Thus we carefully selected that model configuration for our subsequent experiments. Specifically, we have chosen the model architecture that includes four encoding/decoding layers and one bottleneck layer. With this model configuration, we aim to strike a balance between providing a high Dice Score and maintaining a manageable number of parameters. By doing so, we ensure that our proposed model remains efficient in terms of both accuracy and number of parameters.

Chapter 3

Types of Attention Methods

In the domain of medical image segmentation, incorporation of attention mechanisms is crucial for enhancing model performance and reliability. Attention mechanisms help the model to concentrate on the most pertinant parts of the image, improving the localization of critical structures and handling the variability and complexity inherent in medical data. They help extract multi-scale features, reduce false positives and negatives, and provide interpretability by visualizing the model's focus areas. This focused approach improves data efficiency, making better use of limited labeled data, and allows adaptive computation, enhancing both accuracy and efficiency. Thus, attention mechanisms significantly improve the precision and trustworthiness of segmentation models in medical applications.

Here we present various attention techniques utilized in the experimental process. For each attention mechanism, the backbone model is kept as DW U-Net(Table 2.2) with 4 encoding/decoding layers(Table 2.3) and the attention block is incorporated into the bottleneck layer of the model. The following sections will discuss the workings and theoretical foundations of three different attention techniques followed by our proposed model of attention.

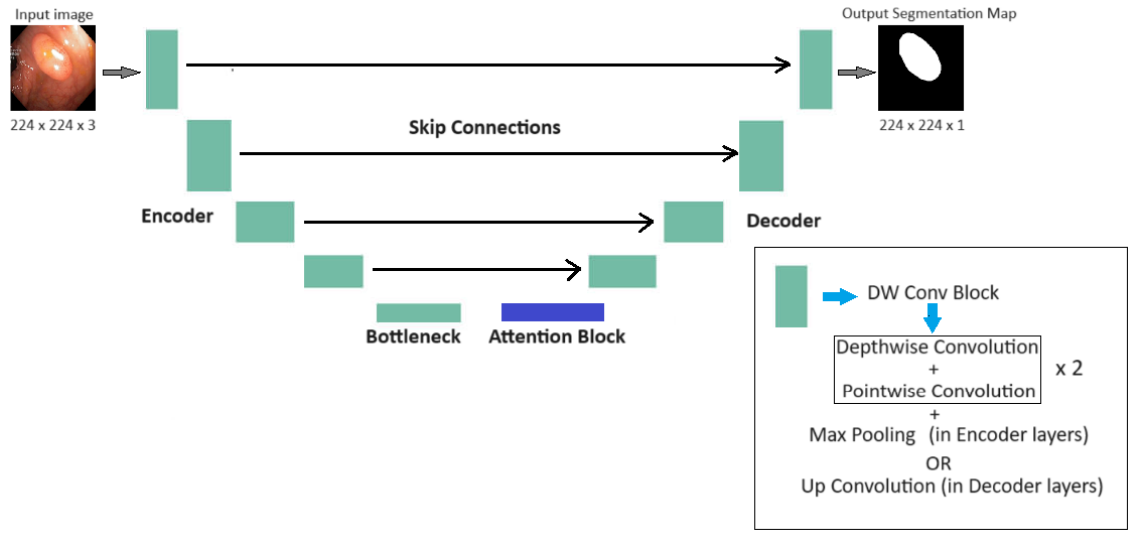

Figure 3.1: Basic Attention based DW U-Net model

3.1 Convolution Block Attention Mechanism(CBAM)

Given a feature map $\mathbf{F} \in \mathbb{R}^{C \times H \times W}$ as input at a certain stage, CBAM sequentially derives a 1D channel attention map $\mathbf{M_c} \in \mathbb{R}^{C \times 1 \times 1}$ and a 2D spatial attention map $\mathbf{M_s} \in \mathbb{R}^{1 \times H \times W}$ as shown in Figure 3.2. The comprehensive attention process can be depicted as[4]:

$$\mathbf{F}' = \mathbf{M}_c(\mathbf{F}) \otimes \mathbf{F}$$
$$\mathbf{F}'' = \mathbf{M}_s(\mathbf{F}') \otimes \mathbf{F}',$$

where $\otimes$ represents element-wise multiplication.

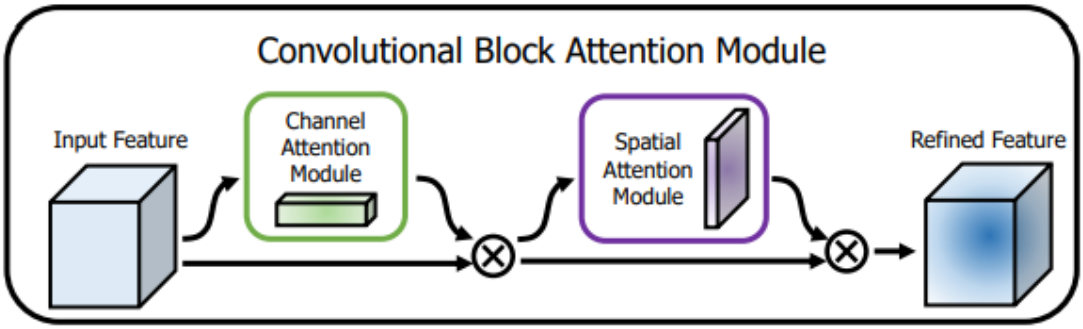

Figure 3.2: An illustration of CBAM module[4]. The module has two sequential subsections: channel attention and spatial attention.

3.1.1 Channel Attention Block

By harnessing the inter-channel relationships within the features, the channel attention map[4] is generated . It identifies 'what' parts of the input image are significant. In this process, the spatial dimension of the feature map is compressed to compute channel attention efficiently. Average pooling and max pooling are utilized to aggregate spatial information and distinguish object features.

Let $\mathbf{F^c_{avg}}$ and $\mathbf{F^c_{max}}$ represent the average-pooled features and max-pooled features, respectively. These feature maps are fed into a shared network to generate the channel attention map $\mathbf{M_c}$. In our case, the shared network consists of two convolutional layers. The first convolutional layer reduces the number of channels by a factor of r and then the second convolutional layer restores the number of channels. After the shared CNN network is applied to each feature map, the output feature vectors are merged using element-wise addition. Precisely, the channel attention [4] is computed as:

$$\begin{aligned}\mathbf{M_c}(\mathbf{F}) &= \sigma(\text{CNN}(\text{AvgPool}(\text{F})) + \text{CNN}(\text{MaxPool}(\text{F}))) \\ &= \sigma(\mathbf{W_2}(\mathbf{W_1}(\mathbf{F^c_{avg}})) + \mathbf{W_2}(\mathbf{W_1}(\mathbf{F^c_{max}}))),\end{aligned}$$

where σ denotes the sigmoid function and $\mathbf{W_1}$, $\mathbf{W_2}$ are the shared CNN weights of the two layers. Figure 3.3 depicts the mechanism of channel attention.

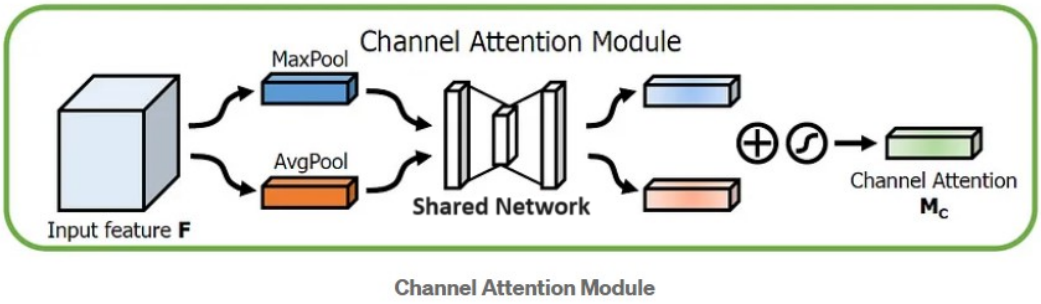

Figure 3.3: Channel Attention Block[4]

3.1.2 Spatial Attention Block

The spatial attention mechanism enhances the channel attention by identifying the specific locations of informative regions within an image or focusing on 'where' the informative parts of an image are. To compute spatial attention, both max pooling and average pooling are applied along the channel axis. The resulting pooled features are joined to form an efficient feature descriptor. A convolutional layer is subsequently applied to this concatenated feature map to produce a spatial attention map $\mathbf{M_s}(\mathbf{F}) \in \mathbb{R}^{\mathbf{H}\times\mathbf{W}}$, which identifies the regions of the image to highlight or diminish.

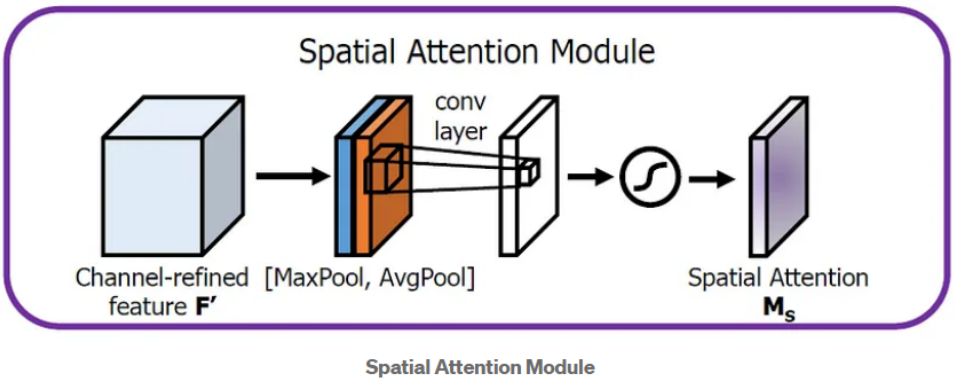

Figure 3.4: Spatial Attention Block[4]

Channel specific information of a feature map is aggregated by using two pooling operations, generating two 2D maps: $\mathbf{F^s_{avg}} \in \mathbb{R}^{\mathbf{1}\times\mathbf{H}\times\mathbf{W}}$ and $\mathbf{F^s_{max}} \in \mathbb{R}^{\mathbf{1}\times\mathbf{H}\times\mathbf{W}}$. Those are then concatenated and convoluted by a standard convolution layer with kernel size 7×7, producing a 2D spatial attention map. Precisely, the spatial attention[4] is computed as:

$$\begin{aligned}\mathbf{M_s}(\mathbf{F}) &= \sigma(f^{7\times7}([\text{AvgPool}(\text{F}); \text{MaxPool}(\text{F})])) \\ &= \sigma(f^{7\times7}([\mathbf{F^s_{avg}}; \mathbf{F^s_{max}}])),\end{aligned}$$

where σ denote sigmoid function and $f^{7\times7}$ denotes convolution with kernel size 7×7. Figure 3.4 depicts the mechanism of spatial attention.

3.2 Self Attention

Attention function converts a query and a group of key-value pairs into an output where the query, keys, values, and output all are represented as vectors. The output is computed by taking a weighted sum of the values. The weights for each value are calculated using a similarity function that assesses the similarity between the query and each key. In this context, the queries and keys have a dimension of d_k, while the values have a dimension of d_v. To compute the weights, we perform the dot product of the queries with all keys, followed by division of each result by $\sqrt{d_k}$, and then application of a softmax function. This process produces the final weights for the values.

In practice, the attention function is computed on a collection of queries at the same time, which are combined into a matrix Q. The values and keys are also stacked together into matrices V and K. Then the matrix of outputs is computed as[5]:

$$\text{Attention}(Q, K, V)[5] = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

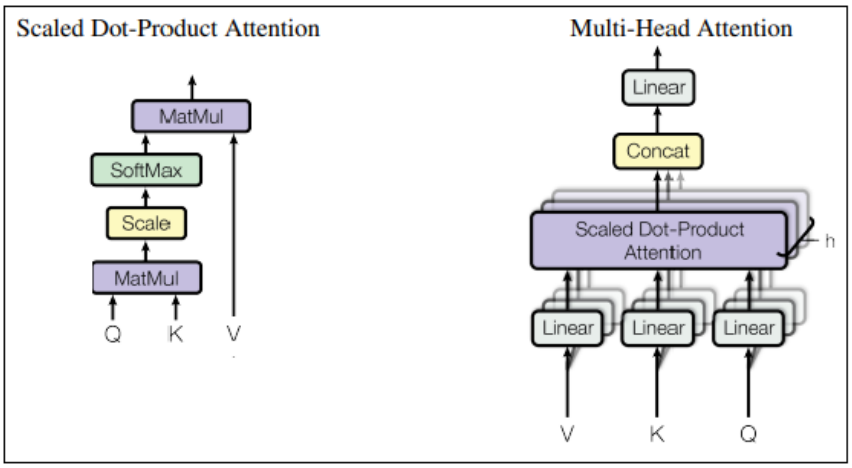

Figure 3.5: (left) Scaled Dot-Product Attention.(right) Multi-head attention containing several attention layers running parallelly.[5]

Until now we saw the entire process confined to only one attention function. Instead of that we can linearly project the queries, keys, and values h times using varying learned linear projections to d_k, d_k, and d_v dimensions, respectively. Now after executing this, the attention function is executed parallelly on each of these projected versions, generating d_v-dimensional output values. These outputs are concatenated and projected again to produce the output values, as illustrated in Figure 3.5. This process is called Multi-Head and is calculated as: .

$$\begin{aligned}\text{MultiHead}(Q, K, V)[5] &= \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \\ \text{where head}_i &= \text{Attention}(QW_i^Q, KW_i^K, VW_i^V),\end{aligned}$$

where $W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$, and $W^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$.

3.3 Convformer

In medical image segmentation, attention collapse within self-attention matrices poses a significant challenge. As neural networks progress through deeper layers, these matrices increasingly become uniform across patches, impeding the capture of essential long-range dependencies within the data. This issue is enhanced by the limited availability of training data in medical imaging tasks. With insufficient data, transformers struggle to learn optimal representations, intensifying the attention collapse problem (Figure 3.6 illustrates this scenario in some attention-based CNN models). Additionally, integrating convolutional neural networks (CNNs) with transformers can introduce biases towards CNN-based representations. This bias usually comes from CNNs' tendency towards relatively smoother convergence, especially when trained on smaller datasets. By addressing attention collapse and balancing the learning dynamics between CNNs and transformers, we can enhance the effectiveness of these models for medical image segmentation tasks.

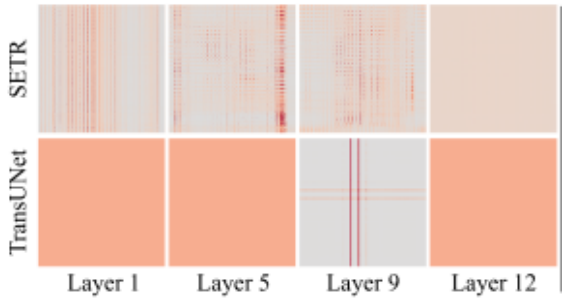

Figure 3.6: Attention Collapse Visualization among layers

To address this issue Convformer[6] was presented. In Convformer, 2D images maintain long-range dependencies without the need for splitting into 1D sequences. It replaces tokenization, self-attention, and feed-forward networks with pooling, CNN-style self attention (CSA), and convolutional feed-forward network (CFFN) respectively. Initially, the image's dimension is reduced through series of convolution and max-pooling operations. CSA adaptively generates convolutional layers to capture appropriate dependencies, adjusting the size for local or global interactions. CFFN further refines pixel features through continuous convolutions.

Vision Transformer[19] being one of the benchmark architectures where attention mechanism is used for image segmentation tasks, a comparison of Convformer's architecture is made with ViT. The Vision Transformer (ViT) is a type of transformer based neural network architecture designed for image classification and segmentation tasks. Contrary to traditional convolutional neural nets (CNNs), which process images as 2D grids of pixels, ViT treats images as sequences of patches. These patches are linearly embedded using positional embedding and then processed by a transformer architecture consisting of self-attention layers and feed-forward layers. On the contrary Convformer processes 2D inputs directly. Convformer employs a pooling module to replace tokenization and preserves both the locality and positional information without the need for any additional positional embeddings. In the core of Convformer, the CNN-style self-attention (CSA)

module takes the place of self-attention (SA) module of ViT to establish long-range dependencies using adaptive and scalable kernels. Additionally, Convformer employs a convolutional feed-forward network (CFFN) to fine-tune pixel features, contrasting with the standard feed-forward network (FFN) in ViT. Unlike ViT, Convformer does not require an upsampling step to resize the output to match the input size as the pooling module adjusts the output size by varying the number of max-pooling operations. Convformer's reliance on convolution operations minimizes the training tensions between transformers and CNNs mentioned earlier.

Pooling vs Tokenization

The pooling module harnesses tokenization's benefits (i.e., adapting the input for transformers in the channel dimension by reshaping and decreasing the input size when necessary) while preserving spatial details, unlike tokenization. For an input $X_{\text{in}} \in \mathbb{R}^{c \times H \times W}$, a convolution with a kernel size of 3×3 is initially applied, followed by batch normalization and ReLU activation to capture local features. To match the resolution corresponding to each patch size S in ViT, a total of $d = \log_2 S$ downsampling operations are performed in the pooling module. Each downsampling operation includes max-pooling with a kernel size of 2×2 and a sequence of 3×3 convolution, batch normalization, and ReLU. As a result, X_{in} is transformed into $X_1 \in \mathbb{R}^{c_m \times \frac{H}{2^d} \times \frac{W}{2^d}}$ through the pooling module, where c_m corresponds to the embedding dimension in ViT. For our scenario, since we incorporate the attention module into the bottleneck layer, we have already passed through four encoding layers. These encoding layers already take care of our requirement of downsampling as each conv block is succeeded by a maxpool operation. Thus this module is not very important in our proposed architectural structure. Following are the main modules that differentiate Convformer from the standard technique of self attention.

CNN-style vs Sequenced Self Attention

Convformer utilizes CNN-style self-attention (CSA) to establish long-range dependencies. CSA dynamically adjusts the receptive field for each pixel by crafting a tailored convolution kernel. In detail, for every pixel $x_{i,j}$ in X_1, the convolution kernel $A^{i,j}$ is formed using two intermediary variables:[6]

$$Q_{i,j} = \sum_{l=-1}^{1} \sum_{g=-1}^{1} E^{q}_{2+l,2+g} x_{i+l,j+g},$$
$$K_{i,j} = \sum_{l=-1}^{1} \sum_{g=-1}^{1} E^{k}_{2+l,2+g} x_{i+l,j+g},$$

where E_q and E_k, belonging to $\mathbb{R}^{c_q \times c_m \times 3 \times 3}$, serve as the learnable projection matrices. Here, c_q denotes the embedding dimension of Q, K, and V, encompassing the features from neighboring pixels within a 3×3 vicinity into $x_{i,j}$. Subsequently, the initial customized

convolutional kernel $I_{i,j}$ for $x_{i,j}$ is determined by evaluating the cosine similarity:[6]

$$I_{i,j}^{m,n} = \frac{\sum_{l=0}^{c_q} Q_{i,j} K_{m,n}}{\sqrt{\sum_{l=0}^{c_q} Q_{i,j}^2} \sqrt{\sum_{l=0}^{c_q} K_{m,n}^2}}.$$

Here, $I_{i,j}^{m,n} \in [-1, 1]$ and it seldom occurs that $I_{i,j}^{m,n} = 0$. The attention score calculation is denoted by $I_{i,j}^{m,n}$. Subsequently, the size of the convolution kernel for $x_{i,j}$ is dynamically adjusted by incorporating a trainable Gaussian distance map M:[6]

$$M_{i,j}^{m,n} = e^{-\frac{(i-m)^2(2d/H)^2 + (j-n)^2(2d/W)^2}{2(\theta \times \alpha)^2}},$$

where $\theta \in (0, 1)$ is also a trainable parameter controlling the receptive field of A, and α is a hyper-parameter regulating the tendency of the receptive field, with θ being proportional to the receptive field. Based on $I_{i,j}$ and $M_{i,j}$, $A_{i,j}$ is computed as $A_{i,j} = I_{i,j} \times M_{i,j}$. This $A_{i,j}$ represents the size-scalable convolutional kernel, which, when multiplied by V (obtained similarly to Q), facilitates the establishment of adaptive long-range dependencies. Finally, a combination of 1×1 convolution, batch normalization, and ReLU activation is employed to integrate features learned from long-range dependencies.

Convolution vs Vanilla Feed-Forward Network

Similar to the working of Feed-Forward layers in the architecture of transformers, the Convolution Feed-Forward Network (CFFN) works with the objective to refine the output generated by the CSA. It contains two modules of CBR, i.e. 1×1 convolution, batch normalization, and ReLU activation. By this modification, CFFN renders Convformer to be entirely CNN-oriented. This design choice circumvents the conflict between CNN and Transformer during training, a scenario frequently encountered in CNN-Transformer hybrid approaches.

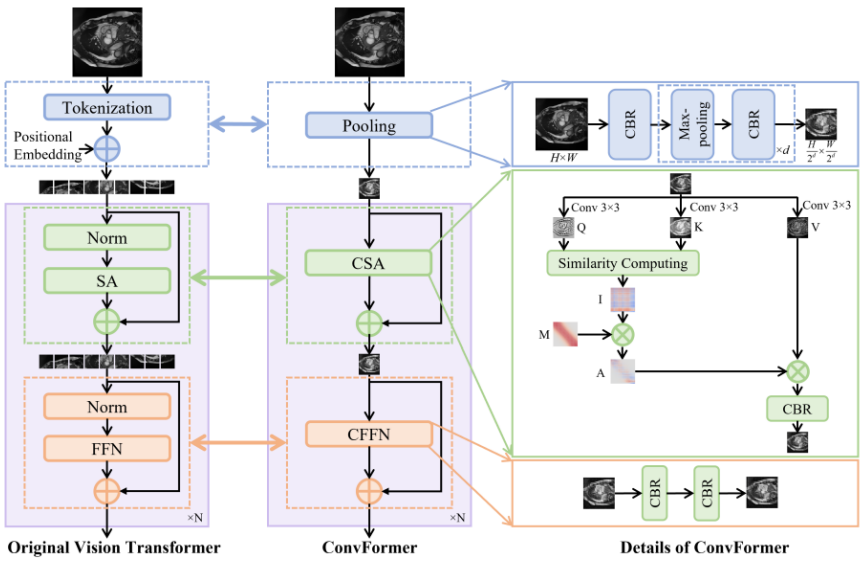

Figure 3.7: Comparison between vanilla Vision Transformer and Convformer. [6]

3.4 Modified Convformer

We have seen that Convformer is presented as an alternative to conventional self-attention techniques. This when plugged in with a CNN architecture avoids the competition between CNN and Transformer during training which often occurs in CNN-Transformer hybrid approaches and overcomes the issues of attention collapse. Having said that, this incurs a load of a huge number of parameters to be trained. This in turn makes its' training a slow process and thus it requires high-performance computational resources like costly GPUs to expedite the training process. To mitigate this issue we have proposed a modified version of the Convformer. We have seen Convformer has 3 major changes compared to standard techniques of self-attention used in ViT. The first one is tokenization and positional embedding is replaced by pooling operations that uphold the necessary spatial information of the image, secondly, the self-attention module is replaced by a convolutional self-attention module whose main objective is to provide a size-scalable convolutional kernel for each pixel and thirdly standard feed-forward network is replaced by Convolutional feed-forward network. As previously mentioned, we do not need to focus on the first change, i.e., replacing tokenization and positional embeddings with pooling. This is because, in our model's inherent architecture, the feature map passes through encoder layers containing a combination of convolution and pooling operations before the attention mechanism is applied. Instead, we focus on the other two changes. Rather than using standard 2D convolutions in CSA and CFFN modules, we used depth-wise and point-wise convolutions in each of them. That is we have used depth-wise convolutions to obtain the values of Q, K, and V in the CSA module, and the CFFN module's convolution network is transformed to DW Convolution Network. Thus our proposed modification of Convformer contains DW-CNN-style self attention and DW-Convolution Feed-Forward Network.

Table 3.1: Parameter Comparison among different Attention based DW U-Net models.

Model	Parameters
DW U-Net with Modified Convformer	18.66 M
DW U-Net with Convformer	68.92 M
DW U-Net with Self Attention	18.6 M
DW U-Net with CBAM	6.12 M

We clearly see from Table 3.1 that number of parameters in Modified Convformer is significantly less(around 4 times) than the number of parameters in the original version of Convformer. In the subsequent sections we will see the performance comparison of these four proposed models on three different types of datasets.

3.5 Experiment and Analysis

To present a comparison study among the proposed models of attention mechanism, we conducted a series of experiments. At first (Experiment 1), we decided on the depth of the transformer and the hidden dimension of the MLP layer of the transformer architecture for the model DW U-Net with Self Attention. After finding out the appropriate depth

and hidden dimension, we used that particular configuration subsequently in DW U-Net with Self Attention for the remaining experiments. Please note that the depth of the transformer in DW U-Net with Modified Convformer and DW U-Net with Convformer has been maintained according to the specifications in [6]. This was done to fully evaluate the model's potential and ensure a fair comparison with our proposed model, DW U-Net with Modified Convformer. Secondly (Experiment 2), we studied the accuracy comparison of the above-mentioned models in Table 3.1 by experimenting on three datasets namely Kvasir-SEG polyp[2], ISIC2017 skin lesion[1] and BRATS brain tumour[23][24][25] datasets. We present in the following subsections a detailed description of these datasets along with experiment details and results of those experiments.

3.5.1 Datasets

Among the three datasets used in this experiment, a description of the Kvasir-SEG polyp[2] dataset is already given in subsection 2.4.1. In this section, we will explore in detail the other two datasets namely ISIC2017 skin lesion[1] and BRATS brain tumour[23][24][25] datasets.

ISIC2017 Skin lesion dataset

The International Skin Imaging Collaboration (ISIC) has evolved into one of the largest and foremost repositories for researchers in the field of machine learning for medical image analysis, particularly in the realm of skin cancer detection and malignancy assessment. Skin cancer stands as the most prevalent of all cancers, with a greater number of diagnoses each year compared to all other cancers combined. In the United States alone, approximately 9,500 new cases are diagnosed daily ([27]). Melanoma, the most lethal form of skin cancer, is projected to reach nearly half a million cases by 2040, marking a 62% increase since 2018. Tragically, one person succumbs to skin cancer every four minutes. Many dermatologists view the escalating incidence of skin cancer as a global epidemic. To enhance survival rates among patients afflicted with skin cancer, early intervention is deemed crucial. Identifying the lesion area is thus of utmost importance, and segmentation activities aid in pinpointing the Region of Interest (ROI) from images. Dermoscopy serves as a widely utilized imaging technique for visualizing the skin surface, yet its diagnostic accuracy heavily relies on the expertise of dermatologists. Consequently, a scarcity of expert resources significantly hampers timely treatment for skin cancers. Given the escalating global incidence of skin cancer, the demand for remote automated diagnosis solutions is growing more critical. This serves as the driving force behind the selection of this dataset for experimental analysis.

The ISIC2017 skin lesion dataset[1] is already pre-splitted at source into the train, test, and validation sets with each having 2000, 600 and 150 images and their corresponding masks respectively. The images are in JPEG format and their corresponding masks are in PNG format. For Experiment 2 we merged the training and validation sets. The main reason behind this merging is that it provides a larger training set which facilitates the model to learn more robust features and improve its capability of generalization. The other

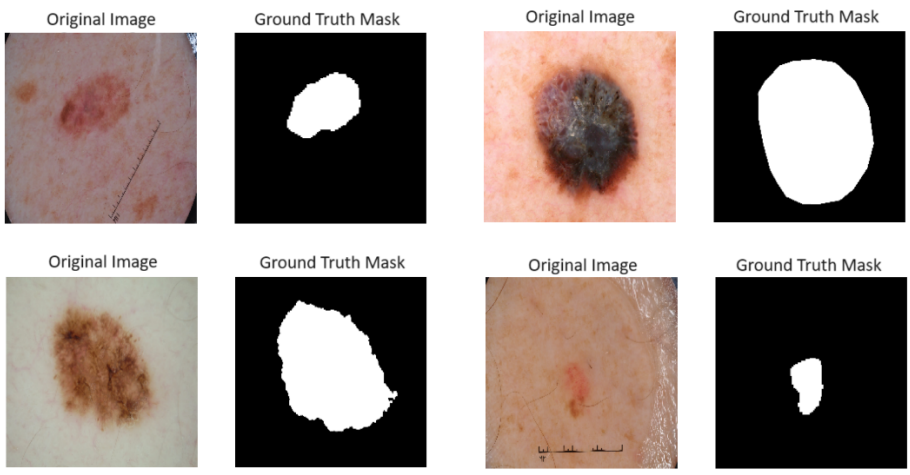

Figure 3.8: Images and their corresponding masks from ISIC2017 skin lesion dataset[1].

point is that the validation set size (150 images) is relatively small compared to the training set (2000 images). This small validation will not provide a sufficiently robust estimate of model performance. By merging the sets, we try to ensure that the model is trained on a more diverse set of examples, potentially leading to better generalization. Also, note that an increase of 150 training samples would not raise any chances of overfitting.

BraTS2020 dataset

Brain tumors are among the most deadly cancers worldwide, classified as either primary or secondary tumors depending on their origin. The predominant histological type of primary brain cancer is glioma, arising from the brain's glial cells and accounting for 80% of all malignant brain tumors. A timely diagnosis is crucial for effective patient treatment. Magnetic Resonance Imaging (MRI) is extensively utilized by radiologists to assess and evaluate brain tumors. It encompasses several complementary 3D MRI modalities, including Fluid-Attenuated Inversion Recovery (FLAIR), T1-weighted, post-contrast T1-weighted (T1ce), and T2-weighted acquired based on different excitation and repetition times.

The training set consists of 369 multi-contrast MRI scans stored as NIfTI files (.nii extension, where NIfTI stands for Neuroimaging Informatics Technology Initiative) with dimensions of 240×240×155. Each scan includes four modalities: Fluid Attenuated Inversion Recovery (FLAIR), native T1-weighted, post-contrast T1-weighted (T1ce), T2-weighted (T2). Each scan is annotated with four classes: background (label 0), GD-enhancing tumor (ET, label 4), peritumoral edema (ED, label 2), and necrotic/non-enhancing tumor core (NET/NCR, label 1). The validation set comprises 125 multi-contrast MRI scans with the same modalities, but the ground truths are concealed. The dataset's objective is to delineate three tumor regions: enhancing tumor, tumor core, and whole tumor area. In our experiment, we evaluated the model's accuracy in identifying the entire tumor area.

In our task, we focused on identifying the whole tumor area. Due to this reason, it was important to binarize the ground truth mask. Thus we first identified a pixel value

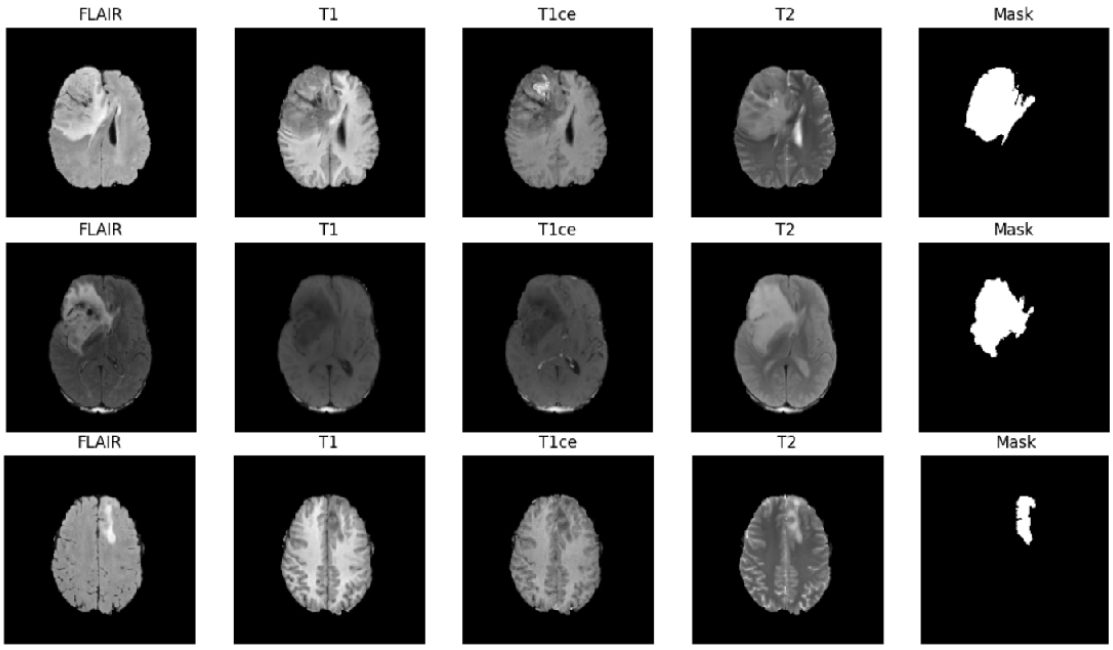

Figure 3.9: Some examples from BraTS2020 dataset. Please note that the masks displayed here are after the processing done on them. We will see in the subsequent parts how the masks are transformed from their original representations.

that was suitable for the threshold and then applied thresholding to binarize the mask. Figure 3.10 shows the original as well as the transformed masks.

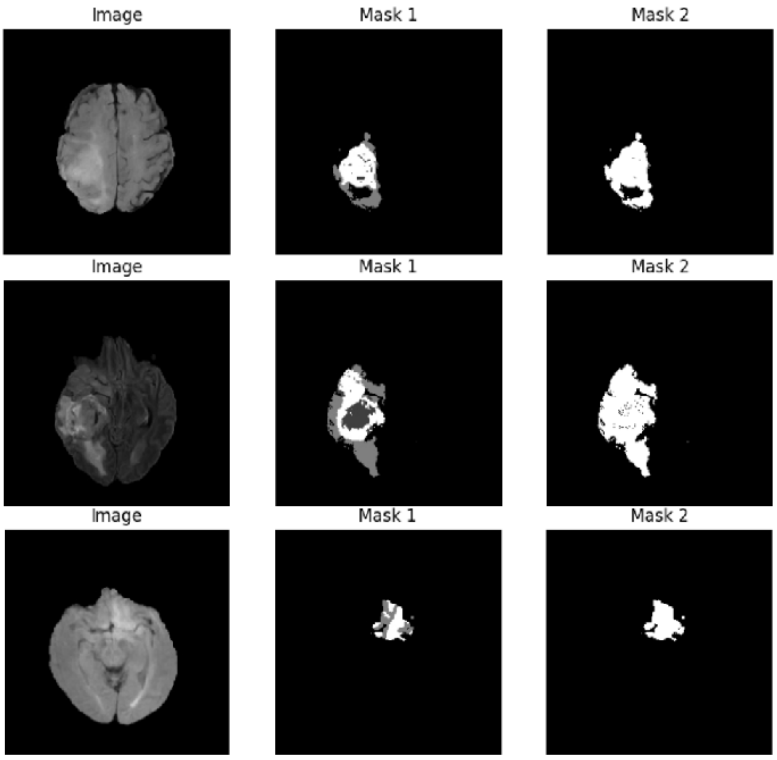

Figure 3.10: Transformation of Masks. Mask1 indicates the ground truth and Mask2 indicates the transformed mask. Please note that the images are based on the FLAIR modality.

After converting the (.nii) files into corresponding JPG format, it was observed that out of the 155 images out of each scan, the first ten and the last ten did not contain any significant information. It was also observed that images at some particular intervals contained redundant information. Thus we chopped off the first ten and last ten images from each scan and considered images at an interval of 9. Other than this we removed the 354th scan due to some technical glitches present in that particular file. Out of 155 images in each scan, we considered only 15 of them. This pre-processing technique helped us decrease the sample size of the image set which was previously 57195 (155×369) to 5520 (15×368). This reduction helped us greatly in saving computational time and also reduced the chances of overfitting.

3.5.2 Experiment Details

For Experiment 1, we used the Kvasir-SEG polyp dataset[2] which contains polyp images and their corresponding ROIs carefully delineated by experienced gastroenterologists and expert medical practitioners. As already mentioned earlier, this dataset contains 1000 images and their corresponding masks which were divided into train and test datasets. For our experiment, we resized the images as well as their masks to dimension 224×224. We used Binary Cross Entropy as the loss function and Dice Score as the evaluation metric(please refer section 2.4.2 and ADAM[26] as the optimizer. The training process was executed for 100 epochs and the dice score was calculated on test dataset.

For Experiment 2, we conducted a comparative study to gauge the effectiveness of various attention techniques described in earlier subsections. The goal was to determine how each technique impacts the performance of our model on the given task of segmentation. For this experiment, we used 3 datasets Kvasir-SEG polyp dataset, the ISIC2017 skin lesion dataset, and the BraTS2020 brain tumor dataset and results are reported for each model on each dataset separately. When we experimented on the Kvasir-SEG polyp dataset, we used five five-fold cross-validation technique. Since this dataset has less number of images, five fold cross validation ensured all data points to be used for both training and testing, maximizing the use of the dataset and providing a comprehensive evaluation. Also, averaging the performance across five different splits reduces the impact of any single split being unrepresentative, leading to more reliable and robust performance estimates. For the ISIC2017 skin lesion dataset and BraTS2020 dataset, we did not perform any cross-validation techniques because, firstly, the ISIC2017 dataset already has its training and test dataset pre-splitted at the source and secondly cross-validation, especially on large medical imaging datasets like BraTS2020, is computationally intensive and time-consuming. Please note that we had to perform a train test split on the BraTS data despite having a validation set because the validation set is hidden at the source and it is not available for general use.

Other than cross-validation, when we experimented on the Kvasir-SEG polyp dataset, we used several transformations like random rotation by 90 degrees, flip and resize to dimensions 224×224 on both the image and its' corresponding mask. Applying transformations such as random rotation and flipping helps to artificially increase the size and

diversity of the training dataset. This makes the model more robust and helps it generalize better to unseen data by exposing it to a wider variety of scenarios. On the other hand, we only used resize to 224×224 transformation when working with ISIC2017 and BraTS2020 datasets. The reason behind it is, that the orientation and position of skin lesions are relatively consistent in clinical images. Therefore, augmentations like rotation may not be necessary, as they might not represent the natural variability in the real-world scenarios of skin lesion presentations. Similarly, MRI scans of brain tumors have a specific anatomical orientation. Rotating these images could lead to anatomically incorrect data, which might confuse the model rather than help it generalize. Hence the choice of augmentation techniques depends on the specific requirements and characteristics of each dataset. For the Kvasir-SEG polyp dataset, augmentations like rotation and flipping enhance variability and generalization. However, for datasets like ISIC2017 and BraTS2020, maintaining clinical and anatomical integrity is crucial, which might lead to a more conservative approach to augmentation, focusing on transformations that preserve the natural characteristics of the data.

Apart from this, we used Binary Cross Entropy(section 2.4.2) as the loss function and Dice Score(section 2.4.2) as the evaluation metric, and ADAM[26] as the optimizer with learning rate 0.001. The training process was executed for 150 epochs and the dice score was calculated on the test dataset.

3.5.3 Results and Discussion

First, we look at the result of the first experiment where did the experiment to find out the depth of the transformer and the hidden dimension of the MLP layer of the transformer architecture. In this experiment, we take some standard depth and hidden dimension values and implement the same in our architecture DW U-Net with Self Attention and find out the optimal combination. In this experiment, we did not use any transformation other than resizing both the image and its' corresponding mask to dimensions 224 × 224 and also ensured that the training conditions were similar for all the models. This was purposefully done such that no external factors could affect the results of the experiments. Below is the result of this experiment.

Table 3.2: Comparison study of different architectures of DW U-Net with Self Attention

Model	Number of parameters	Dice Score
4 encoding/decoding & 1 bottleneck with Self Attention (with depth=1 and mlp_dim=128)	8.36 M	0.6705
4 encoding/decoding & 1 bottleneck with Self Attention (with depth=4 and mlp_dim=512)	18.6 M	0.6816
4 encoding/decoding & 1 bottleneck with Self Attention (with depth=6 and mlp_dim=2048) (tested with the parameters of ViT[19])	43.79 M	0.6238

From Table 3.2, it is clear that the model with 4 encoding/decoding layers and 1 bottleneck layer with Self Attention(depth = 4 and mlp_dim=512) achieves the highest Dice Score with 18.6 million parameters. Increasing the depth and mlp_dim significantly (depth = 6 and mlp_dim = 2048) results in a much larger model with 43.79 million

parameters, but the performance decreases significantly. This clearly indicates that merely increasing the model complexity does not always lead to enhanced performance but can even degrade it due to the potential chances of overfitting. The intermediate model appears to balance the number of parameters and performance effectively, indicating that there is an optimal range of model complexity specific to this task. Too few parameters might limit the model's capacity to learn, while too many parameters might lead to overfitting or other issues.

Thus to summarize, Table 3.2 highlights that there is a balance to be struck between model complexity and performance. A moderately complex model offers the best performance in this comparison, whereas both simpler and much more complex models perform worse. This also highlights the importance of tuning model parameters to find out the optimal model architecture for any given task.

Now we dive into the results of Experiment 2. For Experiment 2, we conducted a comparative study to evaluate the effectiveness of the various described attention techniques. The goal was to determine how each technique impacts the performance of our model on the given task of segmentation. Here we used 3 datasets Kvasir-SEG polyp dataset, the ISIC2017 skin lesion dataset, and the BraTS2020 brain tumor dataset and results are reported for each model on each dataset separately.

Results based on Kvasir-SEG Polyp Dataset

Firstly, we look into the results of Kvasir-SEG polyp dataset. Note that, here we applied five fold cross validation technique and reported the result corresponding to each fold.

Table 3.3: Comparative Study based on Kvasir-SEG polyp data for the different Attention Techniques.

Models	Parameters	Dice Score					Mean Score	Standard Deviation
		K=1	K=2	K=3	K=4	K=5		
DW U-Net with Modified Convformer	18.66 M	0.8383	0.8311	0.8242	0.8107	0.818	0.8245	±0.0108
DW U-Net with Convformer	68.92 M	0.8247	0.8072	0.8118	0.8387	0.7906	0.8146	±0.0182
DW U-Net with Self Attention	18.6 M	0.7863	0.7786	0.75	0.7868	0.7226	0.7649	±0.0280
DW U-Net with CBAM	6.12 M	0.8008	0.7831	0.7536	0.801	0.765	0.7807	±0.0212

From Table 3.3, we observe that for each fold (from K=1 to K=5), the Dice Scores are reported for all four models. Additionally, the average Dice Score and its standard deviation are provided for each model. A closer look at these results reveals that the models DW U-Net with Modified Convformer and DW U-Net with Convformer consistently achieve the highest Dice Scores across all folds. This suggests that, regardless of the training and testing sets, these two models outperform the other models in terms of segmentation accuracy.

We also observe that DW U-Net with Modified Convformer not only has a higher mean Dice Score compared to DW U-Net with Convformer but also exhibits a lower standard deviation. This suggests that DW U-Net with Modified Convformer is not only more accu-

rate but also more consistent in its performance compared to DW U-Net with Convformer. With fewer parameters, DW U-Net with Modified Convformer is more efficient and leads to faster training and inference times. The combination of high accuracy, low variability, and fewer parameters implies that DW U-Net with Modified Convformer is more robust and resource-efficient.

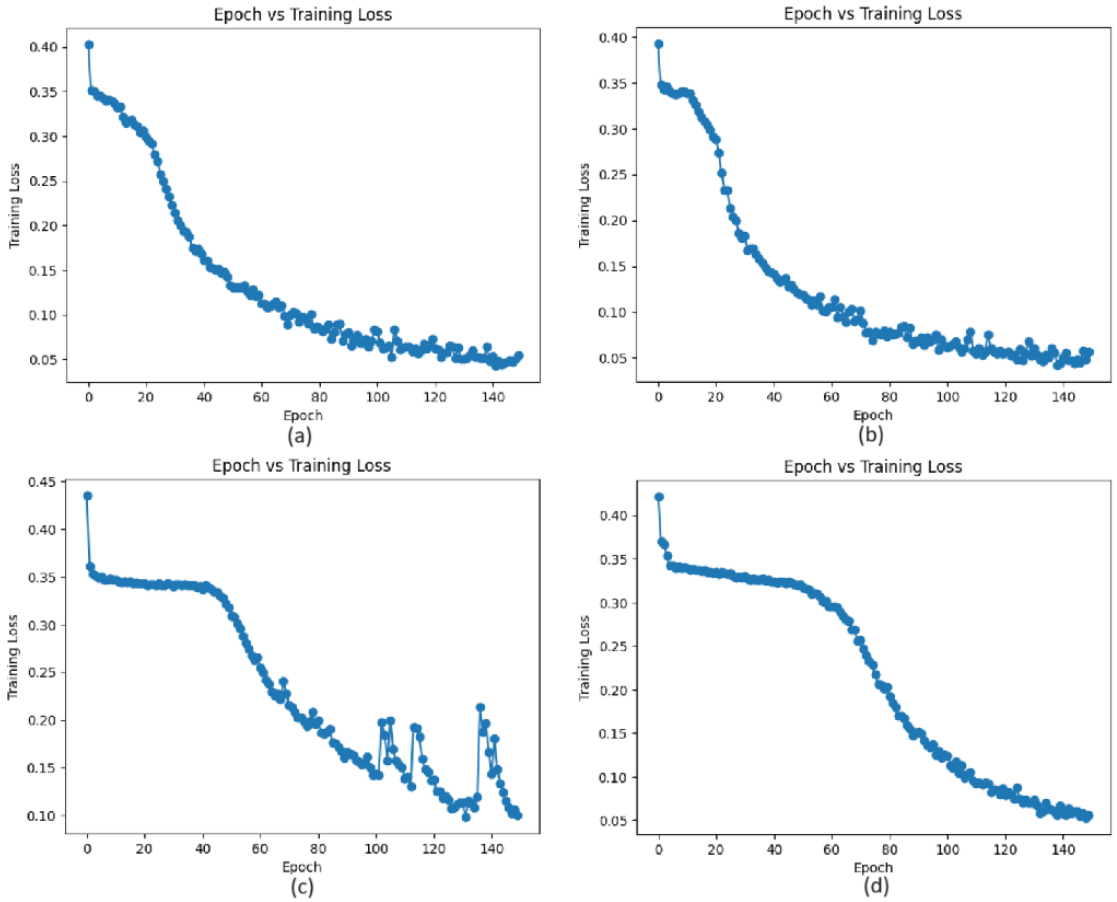

Figure 3.11: Training graphs of the models on Kvasir-SEG polyp dataset. (a) denotes Loss vs Epoch graph for DW U-Net with Modified Convformer. Similarly (b), (c) and (d) indicates to DW U-Net with Convformer, DW U-Net with Self Attention and DW U-Net with CBAM respectively.

From Figure 3.11, it is evident that the training process is most stable for DW U-Net with Modified Convformer, as indicated by its smooth training curve compared to the other models. This smoothness suggests that the model experiences fewer fluctuations in training loss, which can be attributed to better optimization and convergence behavior. Additionally, the stability of the training curve indicates that the model is less prone to overfitting and is able to learn more generalizable features from the data.

From Figure 3.12, it is evident that DW U-Net with Modified Convformer is the best-performing model among the compared models. In the fourth row, DW U-Net with Modified Convformer uniquely identifies the connection between the two tumor areas as shown in the ground truth mask, whereas other models depict the mask as two separate regions. This demonstrates that our model is more adept at learning spatial characteristics compared to others.

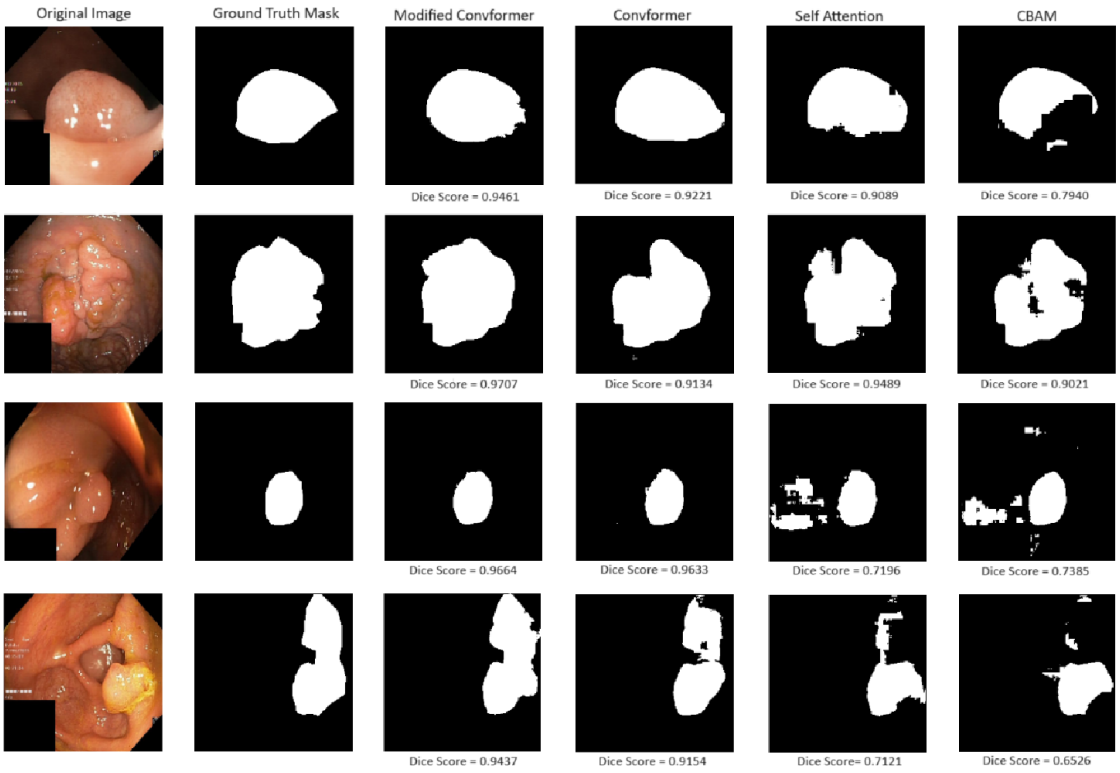

Figure 3.12: Visualization of Results on polyp dataset.

Results based on ISIC2017 Skin Lesion Dataset

Now we look into the results of the ISIC2017 Skin Lesion Dataset. For this dataset, the data was pre-splitted at the source itself. Thus there was no need to manually shuffle and split the data and neither there was any need to apply any cross-validation technique. Below are the results.

Table 3.4: Comparative Study based on ISIC2017 skin lesion dataset for the different Attention Techniques

Model	Parameters	Dice Score
DW U-Net with Modified Convformer	18.66 M	0.8287
DW U-Net with Convformer	68.92 M	0.8199
DW U-Net with Self Attention	18.6 M	0.8023
DW U-Net with CBAM	6.12 M	0.7995

The comparative study presented in Table 3.4 reveals several key insights regarding the performance of different attention techniques on the ISIC2017 skin lesion dataset. The DW U-Net with Modified Convformer achieves the highest Dice Score of 0.8287. This indicates its superior effectiveness in segmenting skin lesions compared to the other models. Additionally, it maintains a relatively moderate parameter count of 18.66 million, suggesting a well-balanced trade-off between model complexity and performance. In contrast, the DW U-Net with CBAM which has the lowest parameter count at 6.12 million, achieves a Dice Score of 0.7995. This makes it the least computationally intensive but slightly less accurate model. Interestingly, the DW U-Net with Convformer, despite having the highest

parameter count at 68.92 million, only reaches a Dice Score of 0.8199, illustrating that an increase in parameters does not necessarily lead to better performance and may result in diminishing returns. The DW U-Net with Self Attention, with 18.6 million parameters, attains a Dice Score of 0.8023, positioning itself as a viable but less optimal technique compared to the Modified Convformer. However, its' training process being very erratic(as shown in Figure 3.13), makes it a less viable architecture to be used. Overall, the DW U-Net with Modified Convformer emerges as the most efficient and effective model for skin lesion segmentation on the ISIC2017 dataset, providing the best Dice Score with a reasonable number of parameters.

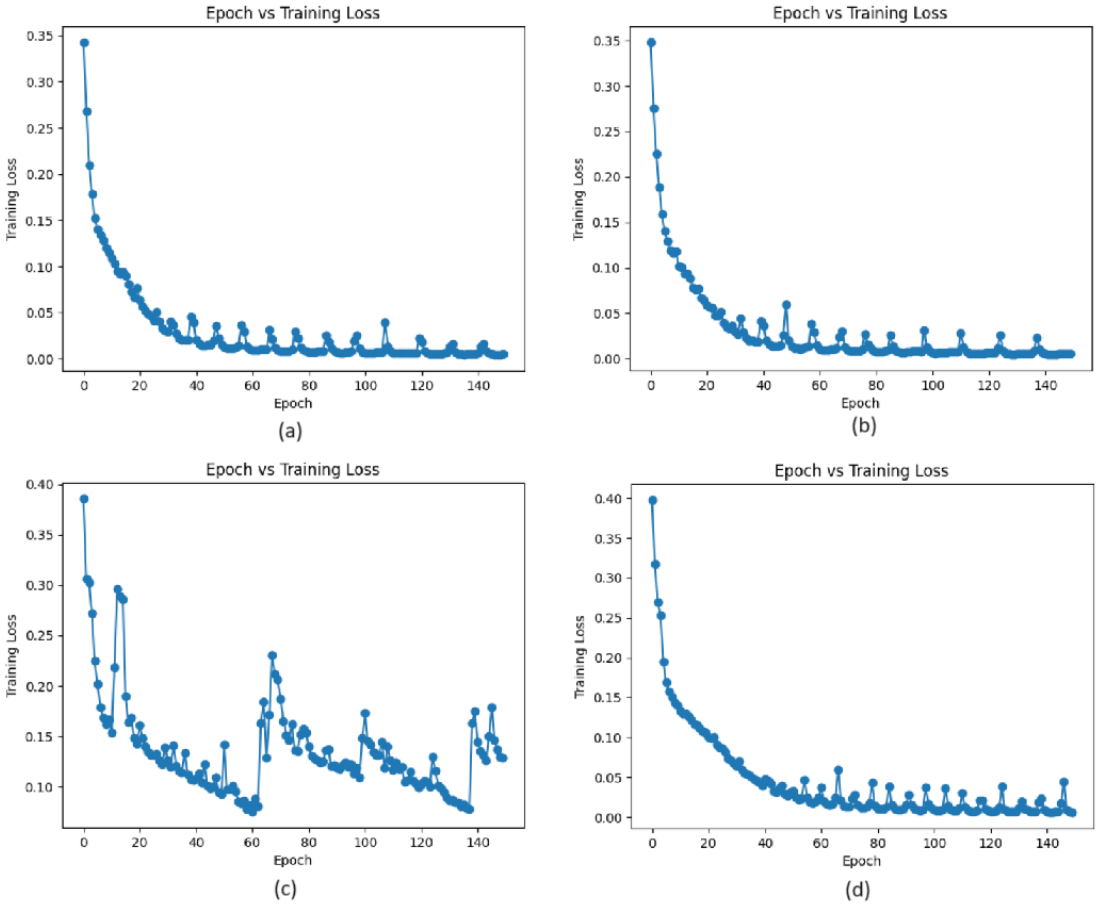

Figure 3.13: Training graphs of the models on ISIC2017 skin lesion dataset. (a) denotes Loss vs Epoch graph for DW U-Net with Modified Convformer. Similarly (b), (c) and (d) indicates to DW U-Net with Convformer, DW U-Net with Self Attention and DW U-Net with CBAM respectively.

From Figure 3.13, it is evident that the training process of all the models apart from DW U-Net with Self Attention is fairly smooth i.e. loss constantly decreases as we progress through the epochs. Now we look into the visualization of this result by checking out some sample test images and see how the models perform on them.

From Figure 3.14, we observe that the performance of the models on most test dataset points is nearly comparable, except for the last image. In this case, DW U-Net with Modified Convformer significantly outperforms the other models, demonstrating nearly

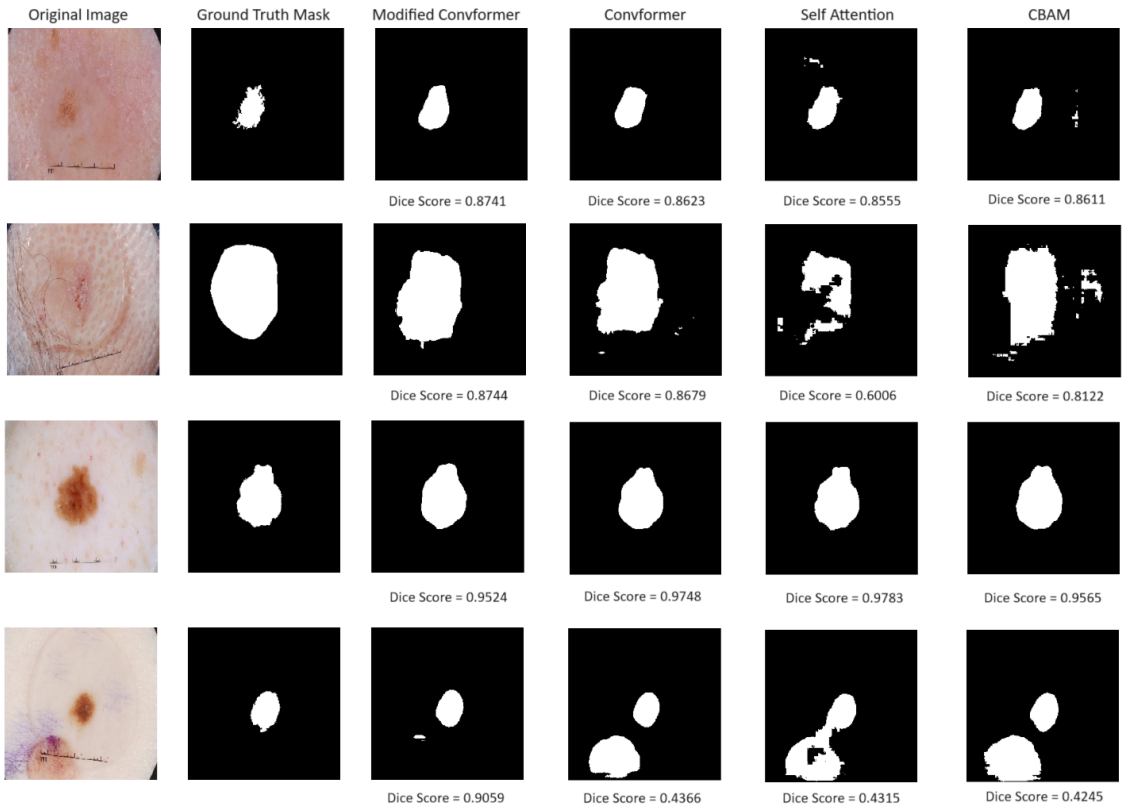

Figure 3.14: Visualization of Results on skin lesion dataset.

twice the performance. This indicates that the Modified Convformer mechanism excels in texture analysis compared to the other models and can identify false positive cases more accurately.

Results based on BraTS2020 Brain Tumour Dataset

Lastly, we look into the results of the BraTS2020 brain tumor dataset. As described in the previous sections section 3.5.1, BraTS2020 is a multimodal dataset. It contains modalities namely FLAIR, T2-weighted, T1-weighted, and post-contrast T1-weighted(T1ce). The training set contained scans of 369 patients where each scan was of dimension $240\times240\times155$. We already discussed earlier about the pre-processing of this data in detail in the previous sections. Despite having a validation set, we needed to split the training dataset into train and test data due to the restricted usage of the validation set. Our goal here is to identify the whole tumor area. For that not only did we compile our models on the different modalities separately but also concatenated all the modalities and tested our models on the concatenated entity (having dimension $224\times224\times4$). Below we provide the results of those experiments.

From Table 3.5, it is evident that the FLAIR and T2 modalities are the most effective in identifying the entire tumor region. The results reveal that different models achieve the highest Dice Score for different MRI modalities, indicating each model's unique strengths in brain tumor segmentation. For the FLAIR modality, the DW U-Net with Modified Convformer achieves the highest Dice Score (0.8508), demonstrating its effectiveness at

Table 3.5: Comparative study based on the BraTS2020 brain tumor dataset for various attention techniques reported across each modality.

Models	Parameters	Dice Score			
		FLAIR	T2	T1	T1ce
DW U-Net with Modified Convformer	18.66 M	0.8508	0.8004	0.7482	0.7242
DW U-Net with Convformer	68.92 M	0.8458	0.8017	0.748	0.7321
DW U-Net with Self Attention	18.6 M	0.8405	0.7661	0.7233	0.6955
DW U-Net with CBAM	6.12 M	0.8387	0.811	0.7268	0.7145

capturing features relevant to detecting tumor regions in FLAIR images, which are known for their high sensitivity to edema. In the T2 modality, the DW U-Net with CBAM attains the highest Dice Score (0.811). Benefiting from CBAM's ability to focus on important spatial and channel-wise features, DW U-Net with CBAM is able to highlight fluid-filled structures and provide good contrast between normal and pathological tissues. For the T1 modality, the DW U-Net with Modified Convformer again stands out with the highest Dice Score (0.7482), leveraging the excellent anatomical detail provided by T1-weighted images for accurate tumor segmentation. Lastly, in the T1ce modality, the DW U-Net with Convformer excels with a Dice Score of 0.7321. T1ce images include contrast enhancement, crucial for highlighting more active or aggressive tumor areas. Convformer's robust feature extraction capabilities outperform the other models in this case. These findings mark the need to consider the specific strengths of different models for each imaging modality in brain tumor segmentation, suggesting that an ensemble approach or modality-specific model selection could yield the best overall performance in clinical practice. Now we present below the result considering all the modalities together.

Table 3.6: Comparative study based on the BraTS2020 brain tumor dataset for various attention techniques, considering all modalities together.

Models	Parameters	Dice Score
DW U-Net with Modified Convformer	18.66 M	86.76
DW U-Net with Convformer	68.92 M	86.74
DW U-Net with Self Attention	18.6 M	86.66
DW U-Net with CBAM	6.12 M	85.65

The Table 3.6 presents a comparative study of different attention techniques based on the BraTS2020 brain tumor dataset, considering all MRI modalities together. Here performance of all the models is relatively close to each other with DW U-Net with Modified Convformer achieving the highest Dice Score of 86.76 with 18.66 million parameters, indicating its superior performance in segmenting brain tumors. Also, we see from Table 3.5 and Table 3.6 that, all the models work better when we consider all the modalities together. This happens because each modality is responsible for identifying a particular area of the tumor, thus concatenating all modalities help us to identify the whole tumor area more accurately. The upcoming sections will provide visualizations of our results, focusing on a single scan. We assessed the performance of our model, DW U-Net with Modified Convformer, based on separate training on each modality and combined training on all modalities. These visualizations will offer insights into how the model performs

better when considering all the modalities together rather than training on each modality separately.

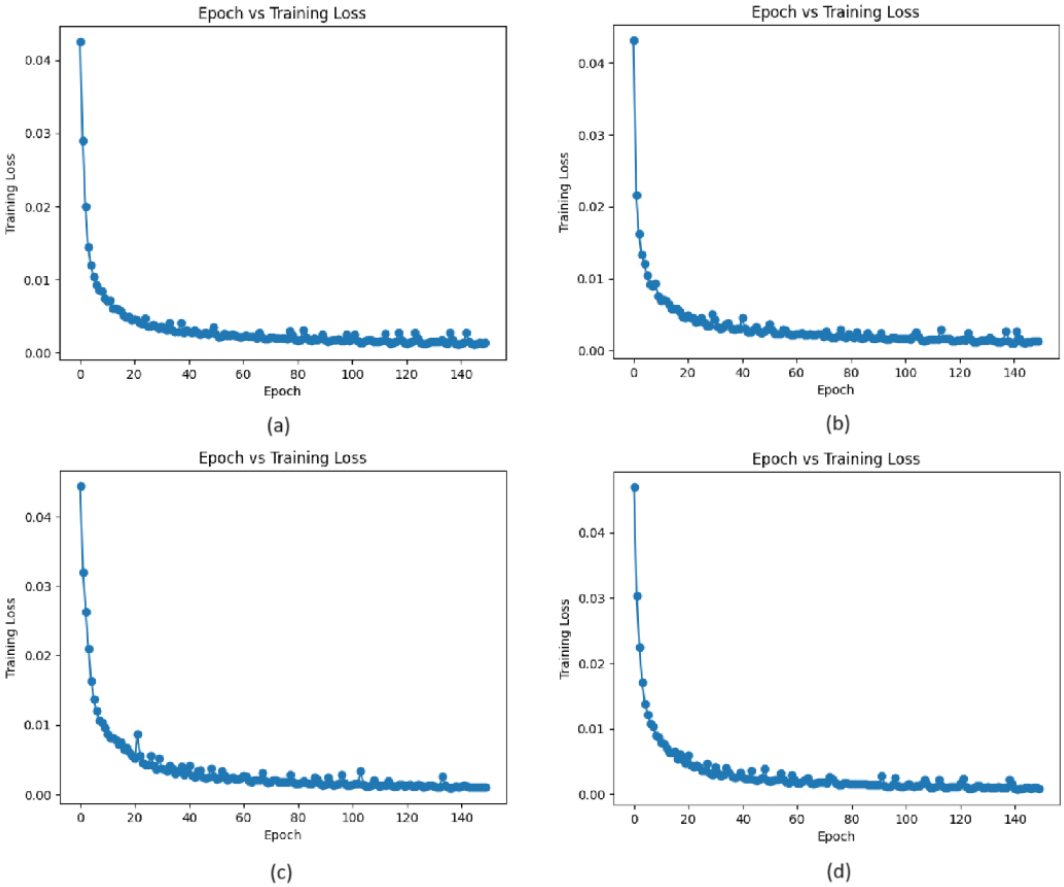

Figure 3.15: Training graphs of the models on BraTS2020 brain tumour dataset. (a) denotes Loss vs Epoch graph for DW U-Net with Modified Convformer. Similarly (b), (c) and (d) indicates to DW U-Net with Convformer, DW U-Net with Self Attention and DW U-Net with CBAM respectively

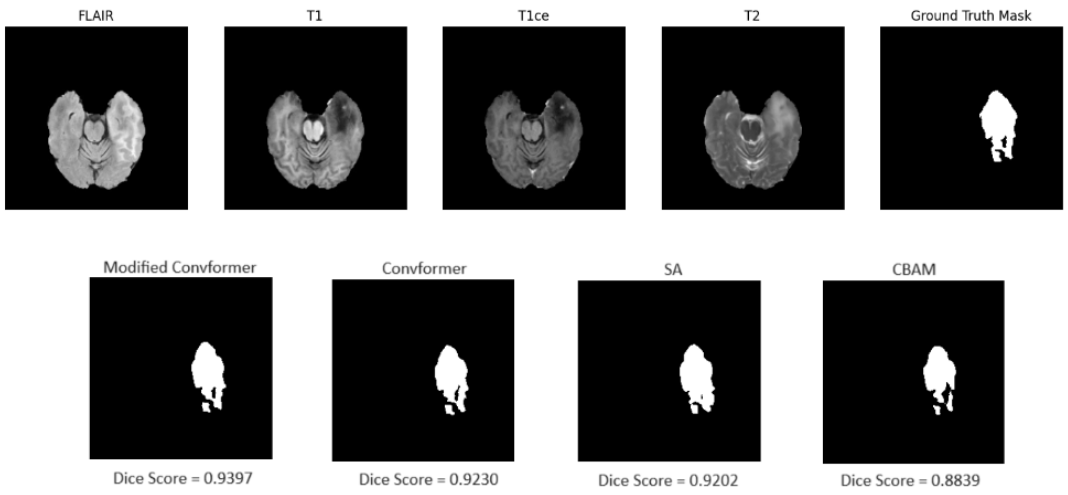

Figure 3.16: Visualization of Results on brain tumour dataset considering all the modalities together.

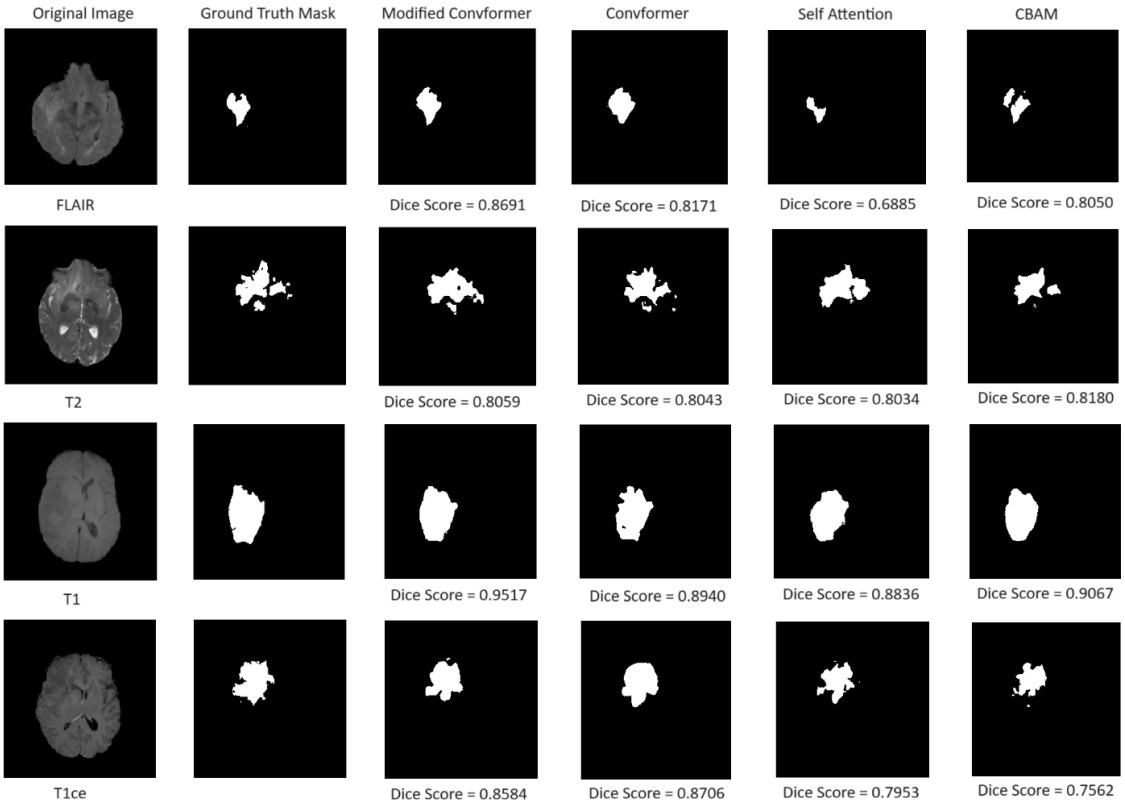

Figure 3.17: Visualization of Results on brain tumour dataset when model was trained on different modalities separately.

Above Figures 3.17 and 3.16 show that when training is done on each modality separately as well as simultaneously, our proposed model DW U-Net with Modified Convformer outperforms other models in both cases.

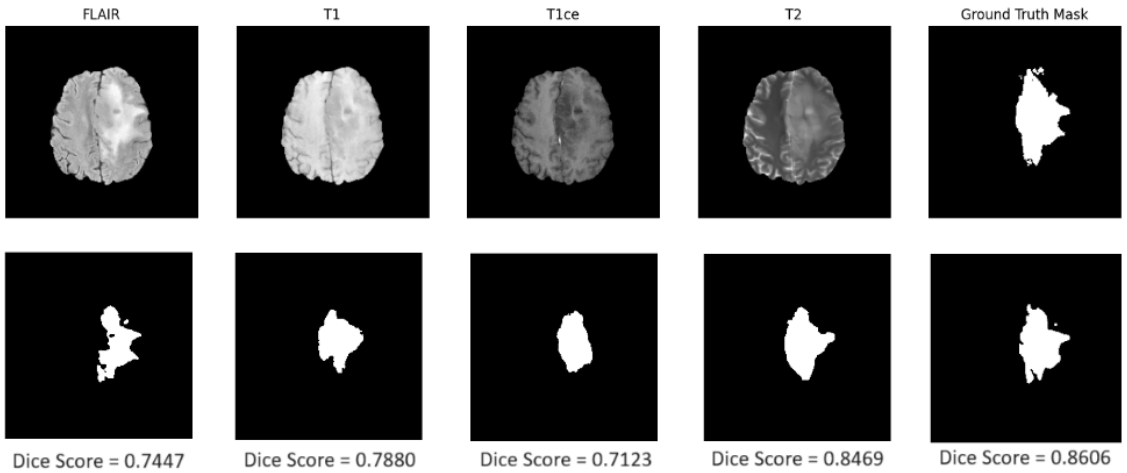

Figure 3.18: This figure illustrates that the DW U-Net with Modified Convformer model achieves higher accuracy when trained on all modalities combined, compared to training on each modality separately. In the second row of the image, the outputs generated from training on FLAIR, T1, T1ce, T2, and their concatenated form are displayed from left to right.

Figure 3.18 illustrates the performance comparison of DW U-Net with Modified Con-

vformer model when trained on different modalities and all modalities combined. The results show that the model achieves higher accuracy when trained on all modalities combined, as opposed to training on each modality separately. This indicates that integrating information from multiple modalities allows the model to better capture the underlying patterns and improve segmentation performance.

From Table 3.3, Table 3.4, Table 3.5, and Table 3.6, it is clear that DW U-Net with Modified Convformer and DW U-Net with Convformer are the top-performing models across all datasets. Among them, DW U-Net with Modified Convformer consistently shows slightly better performance. Additionally, DW U-Net, having fewer parameters, is more computationally efficient.

3.6 State-of-the-art Comparison

For all three datasets namely Kvasir-SEG polyp [2] dataset, ISIC2017 skin lesion[1] dataset and BraTS2020 brain tumor[24][25][23] dataset, we have compared our best proposed model with some of the state-of-the-art models whose results are in this section.

For the Kvasir-SEG polyp dataset, in terms of accuracy, our proposed model performed better than some benchmark models like ResUNet++ and UNet++. It showed comparable results when compared with models like PEFNet, DoubleUnet-DCA, and TransNetR. Below is the table showing the same.

Table 3.7: State-of-the-art Comparison based on Kvasir-SEG polyp dataset

Model	Parameters	Dice Score
Unet++[13]	9.16 M	0.821
ResUNet++[30]	4.06 M	0.8133
DoubleUnet-DCA[28]	30.68 M	0.8516
TransNetR[29]	27.27 M	0.8706
DW Unet with Modified Convformer	18.66 M	0.8245

In this comparative analysis(Table 3.7 of state-of-the-art models' performance on the Kvasir-SEG polyp dataset, it is worthy of mention that the DW Unet with Modified Convformer tries to demonstrate a balance between model complexity and performance. While models like DoubleUnet-DCA and TransNetR achieve higher Dice scores of 0.8516 and 0.8706 respectively, they do so at the expense of significantly higher number of parameter counts (30.68 M and 27.27 M). In contrast, the DW Unet with Modified Convformer, with only 18.66 M parameters, achieves a commendable Dice score of 0.8245.

For the ISIC2017 skin lesion dataset, our proposed model demonstrates comparable performance to leading models such as FAT-Net[31], FAC-Net[32], EIU-Net[33], and DEU-Net[34], which achieve Dice scores of 0.85, 0.8491, 0.8550, and 0.8716, respectively. Our model attains a Dice score of 0.8287, highlighting its competitive efficiency in skin lesion segmentation. Please note that the dice scores of the mentioned models are taken from paper[34].

For the BraTS2020 dataset, we compared our model against those that reported Dice scores on the training data due to restrictions preventing us from evaluating our model

on the validation set. Our proposed model, which achieved a Dice score of 0.8676, outperforms the V-Net with modifications[35], which has a Dice score of 85.13 on training set. Additionally, our model is competitive with other advanced models such as the 3D dResU-Net[36], which has a Dice score of 0.9212 with 30.47 million parameters, and the SGEResU-Net[37], which has a Dice score of 0.9048 with over 19.07 million parameters.

Chapter 4

Conclusion and Future Work

In our work, we have seen comparisons between various kinds of attention techniques like Modified Convformer, Convformer, Self Attention, and CBAM. We used different datasets to gauge the efficiency of those models and validate our findings. We have noticed that our proposed model Modified Convformer outperforms other models across all three datasets namely the Kvasir-SEG polyp dataset, ISIC2017 skin lesion dataset, and BraTS2020 brain tumor dataset in terms of accuracy, and is also computationally efficient. The datasets were meticulously selected from various fields of medical science, including dermatology, gastroenterology, and MRI, to evaluate our model's robustness and adaptability across diverse medical imaging domains. By integrating advanced attention mechanisms and convolution techniques, our model effectively captures intricate patterns and dependencies within the data, leading to significantly improved segmentation performance. Additionally, the Modified Convformer exhibits a balanced trade-off between accuracy and computational cost, making it a viable choice for real-time applications where both precision and speed are critical.

However, there is still room for improvement. Due to time constraints and limited resources, our work remains a preliminary investigation work in this particular domain. Future work could explore several promising avenues to improve our model's performance further. One potential direction is to experiment with different loss functions tailored for medical image segmentation. Using loss functions such as Dice loss, Boundary loss, or focal loss could improve the model's ability to handle imbalanced data and small lesions, leading to more precise segmentation results.

Another area for enhancement is the application of attention mechanisms across various layers of the architecture and not just in the bottleneck layer. By incorporating attention at multiple stages, the model has increased chances of potentially capturing more hierarchical and contextual information, thereby improving its overall performance. Having said that, maintaining computational efficiency while enhancing the model's complexity is also crucial. Thus the use of lightweight attention mechanisms could help in keeping the model efficient without sacrificing performance.

Also as a part of future work, we would like to check our models' performance in multiple-class segmentation problems. Our current study focuses on datasets involving brain tumors, skin lesions, and colorectal polyps. Thus we would like to extend our work

to datasets involving other vital organs such as the heart, lungs, kidneys, and liver. Testing our model's performance on these additional datasets will provide a more detailed and comprehensive understanding of its adaptability and capabilities across a more wider range of medical imaging applications. This expansion will help in developing a more universally applicable and robust model for medical image segmentation, ultimately contributing to improved diagnostic and treatment outcomes in diverse clinical settings. Apart from these points, we would also like to extend this work to weakly-supervised domains. Exploring weakly-supervised learning methods can be beneficial, especially in scenarios where obtaining fully annotated data is challenging or expensive. Weakly-supervised learning allows models to learn from partially labeled or noisy data by leveraging auxiliary information or weak labels.

In summary, Convformer emerges as a superior attention mechanism compared to traditional self-attention techniques. This highlights the significance of integrating convolutions within attention modules, effectively addressing the inherent challenges like attention collapse and other issues encountered in the joint training of CNNs and transformers. Furthermore, the inclusion of Modified Convformer yields significant performance improvements, emphasizing the benefits of utilizing models with reduced parameters to eliminate the risks of overfitting.

While our study demonstrates promising outcomes with Modified Convformer, there exists ample scope for further advancements. Exploring diverse loss functions, extending the application of attention mechanisms across the network, and implementing strategies to ensure computational efficiency are avenues worth exploring. By pursuing these avenues, we can continue to propel the field of medical image segmentation forward, fostering the development of more resilient, precise, and resource-efficient models.

Bibliography

[1] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, *et al.*, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)," in *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pp. 168–172, IEEE, 2018.

[2] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. De Lange, D. Johansen, and H. D. Johansen, "Kvasir-seg: A segmented polyp dataset," in *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26*, pp. 451–462, Springer, 2020.

[3] GeeksforGeeks, "Depth wise separable convolutional neural networks,"

[4] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[6] X. Lin, Z. Yan, X. Deng, C. Zheng, and L. Yu, "Convformer: Plug-and-play cnn-style transformers for improving medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 642–651, Springer, 2023.

[7] M. Sezgin and B. l. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation," *Journal of Electronic imaging*, vol. 13, no. 1, pp. 146–168, 2004.

[8] R. Adams and L. Bischof, "Seeded region growing," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 16, no. 6, pp. 641–647, 1994.

[9] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International journal of computer vision*, vol. 1, no. 4, pp. 321–331, 1988.

[10] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted*

intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18, pp. 234–241, Springer, 2015.

[11] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

[12] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

[13] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pp. 3–11, Springer, 2018.

[14] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu, "Unet 3+: A full-scale connected unet for medical image segmentation," in *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 1055–1059, IEEE, 2020.

[15] J. M. J. Valanarasu and V. M. Patel, "Unext: Mlp-based rapid medical image segmentation network," in *International conference on medical image computing and computer-assisted intervention*, pp. 23–33, Springer, 2022.

[16] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*, pp. 565–571, Ieee, 2016.

[17] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes," *IEEE transactions on medical imaging*, vol. 37, no. 12, pp. 2663–2674, 2018.

[18] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 162, pp. 94–114, 2020.

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[20] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[21] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, "Medical transformer: Gated axial-attention for medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, pp. 36–46, Springer, 2021.

[22] X. Zhao, P. Zhang, F. Song, C. Ma, G. Fan, Y. Sun, Y. Feng, and G. Zhang, "Prior attention network for multi-lesion segmentation in medical images," *IEEE Transactions on Medical Imaging*, vol. 41, no. 12, pp. 3812–3823, 2022.

[23] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, *et al.*, "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE transactions on medical imaging*, vol. 34, no. 10, pp. 1993–2024, 2014.

[24] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos, "Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features," *Scientific data*, vol. 4, no. 1, pp. 1–13, 2017.

[25] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara, C. Berger, S. M. Ha, M. Rozycki, *et al.*, "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge," *arXiv preprint arXiv:1811.02629*, 2018.

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[27] "Skin cancer facts and statistics." `https://www.skincancer.org/skin-cancer-information/skin-cancer-facts`. Last updated: February 2024.

[28] G. C. Ates, P. Mohan, and E. Celik, "Dual cross-attention for medical image segmentation," *Engineering Applications of Artificial Intelligence*, vol. 126, p. 107139, 2023.

[29] D. Jha, N. K. Tomar, V. Sharma, and U. Bagci, "Transnetr: transformer-based residual network for polyp segmentation with multi-center out-of-distribution testing," in *Medical Imaging with Deep Learning*, pp. 1372–1384, PMLR, 2024.

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

[31] H. Wu, S. Chen, G. Chen, W. Wang, B. Lei, and Z. Wen, "Fat-net: Feature adaptive transformers for automated skin lesion segmentation," *Medical image analysis*, vol. 76, p. 102327, 2022.

[32] Y. Dong, L. Wang, S. Cheng, and Y. Li, "Fac-net: Feedback attention network based on context encoder network for skin lesion segmentation," *Sensors*, vol. 21, no. 15, p. 5172, 2021.

[33] Z. Yu, L. Yu, W. Zheng, and S. Wang, "Eiu-net: Enhanced feature extraction and improved skip connections in u-net for skin lesion segmentation," *Computers in Biology and Medicine*, p. 107081, 2023.

[34] A. Karimi, K. Faez, and S. Nazari, "Deu-net: Dual-encoder u-net for automated skin lesion segmentation," *IEEE Access*, vol. 11, pp. 134804–134821, 2023.

[35] L. M. Ballestar and V. Vilaplana, "Brain tumor segmentation using 3d-cnns with uncertainty estimation," *arXiv preprint arXiv:2009.12188*, 2020.

[36] R. Raza, U. I. Bajwa, Y. Mehmood, M. W. Anwar, and M. H. Jamal, "dresu-net: 3d deep residual u-net based brain tumor segmentation from multimodal mri," *Biomedical Signal Processing and Control*, vol. 79, p. 103861, 2023.

[37] D. Liu, N. Sheng, T. He, W. Wang, J. Zhang, and J. Zhang, "Sgeresu-net for brain tumor segmentation," *Math. Biosci. Eng*, vol. 19, pp. 5576–5590, 2022.