



Islamic University of Technology (IUT)

**Medical Image Segmentation Using
Attention-based Residual Double U-Net**

Authors

Akib Mohammed Khan, 170041043

Fahim Shahriar Khan, 170041052

Alif Ashrafee, 170041064

Supervisor

Dr. Md. Hasanul Kabir

Professor, Department of CSE

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

Department of Computer Science and Engineering (CSE)

Islamic University of Technology (IUT)

A Subsidiary organ of the Organization of Islamic Cooperation (OIC)

Academic Year: 2020 - 2021

April, 2022

Declaration of Authorship

This is to certify that the work presented in this thesis is the outcome of the analysis and experiments carried out by under the supervision of Dr. Md. Hasanul Kabir, Professor of the Department of Computer Science and Engineering (CSE), Islamic University of Technology (IUT), Dhaka, Bangladesh. It is also declared that neither this thesis nor any part of this thesis has been submitted anywhere else for any degree or diploma. Information derived from the published and unpublished work of others has been acknowledged in the text and a list of references is given.

Authors:

AKIB.

Akib Mohammed Khan

(Student ID: 170041043)



Fahim Shahriar Khan

(Student ID: 170041052)

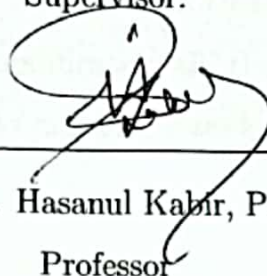


Alif Ashrafee

(Student ID: 170041064)

Approved By:

Supervisor:



A handwritten signature in black ink, consisting of a large, stylized 'H' and 'K' with a long, sweeping underline that extends to the right and then loops back down to the left.

Dr. Md. Hasanul Kabir, Ph.D.

Professor

Department of Computer Science and Engineering (CSE)

Islamic University of Technology (IUT)

Acknowledgement

We would like to thank the Almighty Allah for being able to complete our undergraduate thesis on a good note. We are indebted to **Professor Dr. Md. Hasanul Kabir**, Department of Computer Science & Engineering, IUT for being our adviser and mentor. His motivation, suggestions and insights for this research have been invaluable. Without his support and proper guidance this research would never have been possible. His valuable opinion, time and input provided throughout the thesis work, from first phase of thesis topics introduction, subject selection, proposing algorithm, modification till the project implementation and finalization which helped us to do our thesis work in proper way. We are really grateful to him.

We would also like to thank the head of department, Professor Dr. Abu Raihan Mostofa Kamal for creating a research-friendly environment at IUT. Such a research-friendly environment is vital for proper research.

Abstract

Keywords: Segmentation, Attention Gate, Residual Block, U-Net, Double U-Net.

A common use case for image segmentation in medical-image-based diagnosis is to help clinicians to focus on a specific area of the disease. Manually inspecting polyps from colonoscopy for colorectal cancer or performing a biopsy on skin lesions for skin cancer are time-consuming, laborious, and complex procedures. Automatic medical image segmentation aims to expedite this diagnosis process. The accuracy of image segmentation has increased due to advancements in machine learning techniques and deep learning models. However, there is still room for improvement as there exist various challenges due to the large variation in the appearance of objects in different sizes with no distinct boundaries. To address these issues, we propose a novel-attention based residual Double U-Net architecture that improves on the currently existing skin lesion segmentation networks. We incorporate attention gates on the skip connections and residual connections in the convolutional blocks of Double U-Net, a state-of-the-art (sota) segmentation network. The attention gates allow the model to retain more relevant spatial information by suppressing irrelevant feature representation from the down-sampling path. At the same time, residual connections help to train deeper models by ensuring better gradient flow. We conducted experiments on three datasets: ISIC 2018 (skin lesion), CVC Clinic-DB (polyp), and the 2018 Data Science Bowl (nuclei) datasets and achieved Dice Coefficient (DSC) scores of **91.64%**, **94.35%** and **92.45%** respectively. Further improvement can be achieved by simplifying the structure of our architecture in order to reduce the number of parameters.

CONTENTS

1	Introduction	6
1.1	Motivation	6
1.2	Overview	7
1.3	Problem Statement	8
1.4	Research Challenges	10
1.5	Contributions	11
1.6	Organization of Thesis	11
2	Literature Review	12
2.1	Convolutional Networks for Biomedical Image Segmentation: U-Net	13
2.2	Attention U-Net	14
2.3	Double U-Net	17
2.4	Skin Lesion Segmentation using Generative Adversarial Networks .	19
2.5	Transformer-based model for medical image segmentation	20
2.6	Polar image transformations to improve medical image segmentation	22
3	Proposed Methodology	25
3.1	Overview of our proposed architecture	26
3.2	Data preprocessing: Color Constancy	27
3.3	Squeeze and Excite Block	29

3.4	Atrous Spatial Pyramid Pooling (ASPP)	29
3.5	Attention Gates	30
3.6	Residual Connections	32
3.7	Improvement of our proposed convolutional block over that of Double U-Net	33
3.8	Architecture Variants	33
3.9	Loss Function	34
4	Result Analysis	36
4.1	Dataset Description	36
4.1.1	ISIC 2018	36
4.1.2	CVC-ClinicDB	36
4.1.3	2018 Data Science Bowl	37
4.2	Experimental Setup	39
4.3	Evaluation Metrics	40
4.4	Quantitative Results	40
4.4.1	Ablation study	40
4.4.2	Training progression	42
4.4.3	Comparitive analysis with current state-of-the-art architec- tures	43
4.5	Qualitative Results	46
5	Conclusion and Future Work	49

LIST OF FIGURES

1.1	Examples of semantic segmentation	9
1.2	An example of skin lesion segmentation	9
2.1	U-Net Architecture	14
2.2	Attention U-Net Architecture	16
2.3	Block diagram of the Double U-Net Architecture	18
2.4	Qualitative results for Double U-Net	19
2.5	Architecture of UNet-Critic Model	20
2.6	Architecture of Boundary Aware Network model	21
2.7	Architecture of TransUnet	22
2.8	Pipelines for Image Segmentation Using Polar Transformation	24
3.1	Block diagram of the proposed model architecture	27
3.2	Effect of Color Constancy on RGB color channels	28
3.3	Effect of Shades of Gray algorithm on lesion images	29
3.4	Squeeze and Excite Block Architecture	30
3.5	Demonstration of ASPP	31
3.6	Attention Gate structure	31
3.7	Comparison of convolutional blocks	34
3.8	Block diagram of the Half-Attention Double U-Net	35

3.9	Block diagram of the Full-Attention Double U-Net	35
4.1	ISIC 2018 Dataset examples	37
4.2	CVC-ClinicDB Dataset examples	38
4.3	2018 Data Science Bowl Dataset examples	38
4.4	ISIC 2018 Training Progression	43
4.6	2018 Data Science Bowl Training Progression	43
4.5	CVC-Clinic-DB Training Progression	44
4.7	Qualitative results of our model	48

LIST OF TABLES

2.1	Comparative Analysis of Attention U-Net and U-Net on TCIA Pancreas-CT Dataset	16
4.1	Training hyperparameters for each dataset	39
4.2	Result of our proposed approaches on ISIC 2018 dataset	41
4.3	Comparative result of our proposed model with and without a critic network ISIC 2018 dataset	41
4.4	Results of our proposed model on the effect of adding color constancy and residual connections on ISIC 2018 dataset	42
4.5	Comparative result of our model against the current state-of-the-art models on the ISIC 2018 dataset	45
4.6	Comparative result of our model against the current state-of-the-art models on the CVC Clinic-DB dataset	45
4.7	Comparative result of our model against the current state-of-the-art models on the 2018 Data Science Bowl dataset	46
4.8	Overall improvement on the three benchmark datasets over standalone Double U-Net	46

CHAPTER 1

INTRODUCTION

1.1 Motivation

Skin cancer is one of the most rapidly growing cancer worldwide. Melanoma, a type of skin cancer, accounts for 75% of skin cancer deaths [44], so the proper diagnosis of melanoma is of growing importance. When detected early, the 5-year survival rate for melanoma is 95% [40]. Manually inspecting and performing a biopsy on skin lesions for cancer is time-consuming, laborious, and requires complex clinical experiences and there could be errors due to fatigue. So diagnosing melanoma quickly and accurately is a critical task. This is where computer-aided approaches could be helpful in efficiently diagnosing melanoma from dermoscopic images.

Colorectal cancer is a common form of cancer affecting the colon and rectum. Colonoscopy is a recognized polyp detection method for the early detection and prevention of colorectal cancer. Patients with missed polyps who are diagnosed with advanced-stage colorectal cancer have less than 10% survival rate whereas early-stage diagnosis can ensure survival rates greater than 90% [4]. However, clinicians need to carefully examine the colonoscopy input for polyps which can be challenging due to their various morphological features and sizes. If there are flat polyps, they align with the walls of the rectum, making it indistinguishable from

its surroundings. Automatically segmenting out the polyps can help to facilitate better diagnosis as it can pick up on the slight pixel variations in order to extract out the region of interest, which may not be perceivable to the human eye. [19, 32].

1.2 Overview

Image segmentation is the process of assigning a label to every pixel in an image such that pixels with the same label share similar characteristics. Labeling the pixels in this way results in the region of interest (ROI) sharing the same label. One of the pivotal applications of image segmentation is localizing diagnostically important anatomical structures on medical images, also known as biomedical image segmentation. It is a crucial task for the automation of medical image-based diagnosis. It can be used to identify single structures of an elliptical shape. Heart, kidney cells, skin lesions, polyps, etc. all fall under this category [3]. Segmentation of these structures is one of the important pre-processing steps for other medical tasks like classification or detection.

In the early days, medical segmentation was done using traditional machine learning methods. Celebi et. al. [9] performed unsupervised methods to segment out skin lesions using clustering algorithms. Wong et. al. [49] implemented a stochastic region merging approach on a pixel level and a region level for extracting the lesions from macroscopic images. For automatic polyp segmentation, Gross et. al. [23] used a template matching algorithm where they applied multi-scale filtering for edge detection, the result of which are then compared to a set of elliptic templates. With the advancement in technology and rapid increase in computational resources, however various deep learning models have emerged for the analysis of medical images. Although the results of these models have been impressive, analysis of medical images with lesions, polyps, and other abnormalities with these techniques still experiences some challenges due to the unique and complex features of the skin lesion images. Convolutional Neural Networks (CNNs) have recently had excellent performance across several medical segmentation bench-

marks [34]. One of the most used neural network architectures for biomedical image segmentation is U-Net [41]. It uses a series of CNNs in the encoding path for spatial information and a similar series of CNNs in the contracting path with skip connections to retain contextual information. This allows the network to simultaneously learn context and precise localization. Various modifications and improvements to the U-Net have been proposed. Zhou et. al. [51] introduced U-Net++ where the encoder and decoder are connected via convolutional networks instead of simple concatenation. Oktay et. al. [39] proposed attention gates on the skip connections that allow the model to automatically learn to focus on target regions having irregular and non-standard shapes. Jha et. al. [29] proposed an architecture called Double U-Net based on two U-Nets stacked on top of each other. More recent image segmentation architectures include the use of transformers that help to model long-range contextual information that lead to better performance as seen in DS-TransUnet architecture by Lin et. al. [33] and Boundary Aware Transformer architecture by Wang et. al. [46]. Srivastava et. al. [43] introduced a residual fusion network having the ability to exchange multi-scale features that ensure better propagation of high and low-level features.

1.3 Problem Statement

The goal of this thesis is to delineate the region of interest (skin lesion, polyp, or nuclei) from biomedical images. So taking the images as input, we need to generate a binary mask of the corresponding image where the pixels covering the region of the lesion belong to one class (region of interest) and all the other pixels belong to another class (background).

Medical Image Segmentation can be defined as an automatic process to detect boundaries or region of interest (ROI) within a 2D or 3D image. For segmentation, we assign each pixel a specific class, where the ROI belongs to one class and all the other pixels belong to another class.

Segmentation can be divided into two types:

- **Binary segmentation:** The ROI belongs to one class (1) and all the other pixels belong to the background class (0) as shown in Figure 1.2.
- **Semantic segmentation:** The ROI belongs to one of multiple classes along with the non ROI regions belonging to the background class as shown in Figure 1.1.

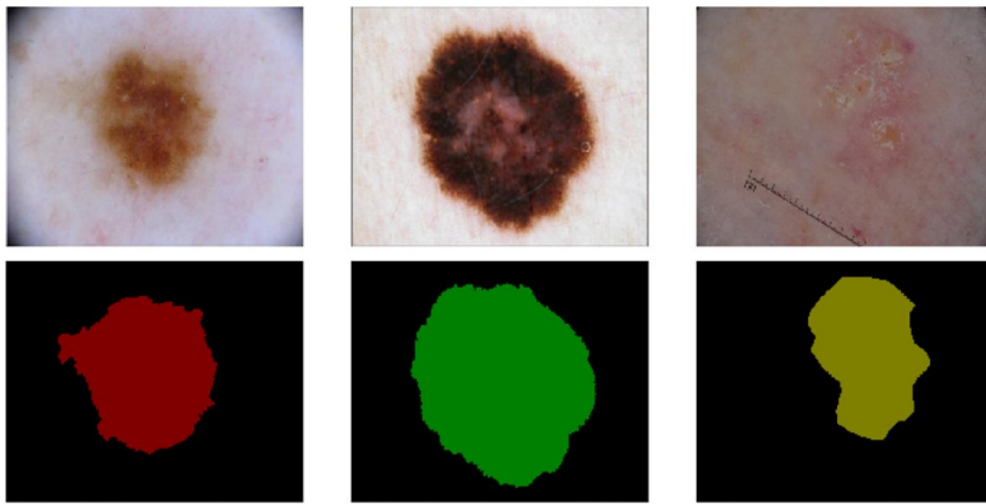


Figure 1.1: Examples of semantic segmentation (Courtesy of [22])



(a) Dermoscopic image containing skin lesion



(b) Corresponding binary mask highlighting the region of interest

Figure 1.2: An example of input skin lesion image and its corresponding output mask (Courtesy of [15])

For example, skin lesions are areas of skin that look different from the surrounding area. They are often bumps or patches, and many issues can cause them. So the task for skin lesion segmentation is to identify the pixels that belong to the ROI in the lesion image and group them together, whether it be binary or semantic.

1.4 Research Challenges

In order to sufficiently train deep models, a large number of training samples are necessary. However, the availability of medical images is very scarce, especially for rare diseases such as melanoma. This poses a significant challenge in the domain of medical image segmentation [30].

Due to huge variations in human skin color, skin lesion segmentation becomes an incredibly complex task. In addition to that, variability in lesion location, size, and shape makes the task even more challenging. Lesions with irregular and fuzzy boundaries also contribute to difficulties in localizing them. The presence of noise and various artifacts like hair, air bubbles, and blood vessels also affect image interpretation by computer-aided lesion segmentation techniques. In some cases, color illumination and low contrast of images where lesions are visually inseparable from the background skin color, pose additional difficulty in segmenting the lesion accurately and may require some form of pre-processing [36, 1].

Polyps in colonoscopy images are also subject to variable sizes and shapes. Moreover, they can be very small and flat, becoming indistinguishable from the mucosa of the colon. Nuclei can also appear densely clustered, making it difficult to extract the overlapped objects.

Apart from issues related to the input images, some of the corresponding ground truth masks acquired from the benchmark datasets also contain noise and are mislabelled. Finally, training our models requires a huge amount of computational resources which poses an additional challenge in our research.

1.5 Contributions

In this report, we have introduced a novel architecture that is built upon the Double U-Net model with our contributions as follows:

- We incorporate **attention gates** that allow the model to retain more relevant spatial information by suppressing irrelevant feature representation from the down-sampling path of the encoding network.
- We include **residual connections** which help to train deeper models by ensuring better gradient flow.
- We apply **color constancy (CC)** as a pre-processing technique that allows the model to give state-of-the-art performance even with less number of data augmentations over the standalone Double U-Net architecture.

1.6 Organization of Thesis

The chapter 2 of our thesis discusses related works in the domain. Chapter 3 gives an overview of the benchmark datasets used for experiments. In chapter 4, we describe our proposed model. The 5th chapter gives a qualitative and quantitative analysis of the outputs of our work. Finally, the 6th and last chapter concludes the work with a brief summary and direction for future research endeavors.

CHAPTER 2

LITERATURE REVIEW

With the dawn of deep learning came a very popular method for image segmentation using CNNs. U-Nets, Attention U-Nets and Double U-Nets are among the most successful approaches toward biomedical image segmentation. Moreover, in recent times, many unique and interesting research directions have been adopted to improve the quality of medical image segmentation. Generative Adversarial Networks [21] have been a popular choice for lesion segmentation purposes. They contain a pair of neural networks, a generator, and a discriminator. The goal of the generator is to produce the segmentation masks while the discriminator tries to differentiate between the original and generated masks. Using the feedback from the discriminator, the generator tries to produce better segmentation results. The arrival of transformer-based architectures [45] in Natural Language Processing (NLP) paved the way for Vision Transformers (ViT) [17] as well for image classification, detection, segmentation, etc. ViTs add positional embeddings to each patch of an image to form an input sequence that is fed to the transformer network which uses a multi-headed attention block so that the image can retain its positional information which is lost when using traditional neural networks. Some preprocessing techniques have also been proposed to improve segmentation tasks. Marin et. al. [3] suggest training on polar transformed images to give

better segmentation results.

2.1 Convolutional Networks for Biomedical Image Segmentation: U-Net

Deep learning has been widely used for medical image segmentation and has become an important research direction in the field of computer vision. The promising ability of deep learning approaches has put them as a primary option for image segmentation, and in particular for medical image segmentation. Especially in the previous few years, image segmentation based on deep learning techniques has received vast attention. U-Net [41] is an architecture for semantic segmentation. As shown in Figure 2.1, it consists of an encoding path followed by a decoding path. The encoding path follows the typical architecture of a convolutional network that properly encodes contextual information. It consists of the repeated application of two 3x3 convolutions (unpadded), each followed by a rectified linear unit (ReLU) and a 2x2 max pooling operation with stride 2 for down-sampling. At each down-sampling step, we double the number of feature channels. Every step in the decoding path consists of an upsampling of the feature map followed by a 2x2 convolution (“up-convolution”) that halves the number of feature channels, a concatenation with the correspondingly cropped feature map from the encoding path, and two 3x3 convolutions, each followed by a ReLU. The decoder path uses the encoded context for accurate localization. Cropping is necessary due to the loss of border pixels in every convolution. At the final layer, a 1x1 convolution is used to map each 64-component feature vector to the desired number of classes. In total the network has 23 convolutional layers. Furthermore, U-Net also has skip connections between encoder and decoder blocks so that spatial information can propagate deep into the network.

U-Net was applied to the segmentation of neuronal structures from electron microscopic readings. The dataset was collected from the EM segmentation challenge [8], on which they achieved a warping error [28] of **0.000353** surpassing exist-

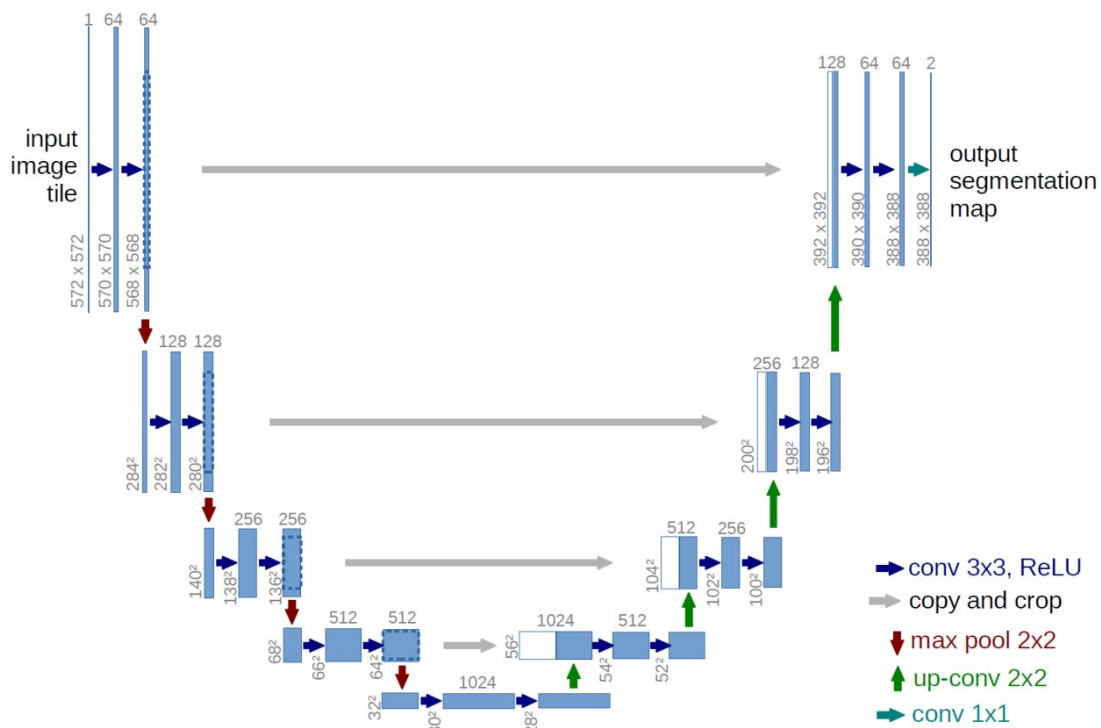


Figure 2.1: U-Net Architecture (Courtesy of [41])

ing methods. More experiments were performed on the segmentation of cells in light microscopic images. They achieved an Intersection over Union (IOU) score of **92.03%** on the ISBI cell tracking challenge 2014 dataset [35], consistently improving upon existing state-of-the-art models at the time.

U-Net consists of a very simple structure of convolutional networks. This poses a range of limitations. The model cannot localize complex features with non-standard shapes and irregular boundaries. The skip connections between the encoder and decoder cannot suppress irrelevant features from the down-sampling layers, for which the model loses the relevant spatial information for proper localization.

2.2 Attention U-Net

Since the skip connections in U-Net are unable to suppress irrelevant features from the down-sampling path, Oktay et. al. [39] suggested a novel Attention Gate (AG)

module on the skip connections. Attention gates learn how to focus more on target regions of various shapes and structures, rather than having equal focus on the entire image. Models trained with AGs know how to suppress irrelevant regions in the image and focus more on extracting important features required for the specific task.

Attention can be of two types:

- **Hard Attention:** The function of attention is to somehow focus on relevant regions more. One way of implementing this would hard attention which crops out the image in places of interest. Since hard attention considers one region at a time to look at so it is not differentiable and hence cannot be learned by backpropagation.
- **Soft Attention:** Weighs different parts of the image. A small weight is given to the regions of lesser importance and a larger weight is given to the regions of more importance. As training progresses, the model learns even more about how to focus on relevant regions.

Soft Attention thus removes the need for explicit image localization modules of cascaded CNNs and also the need for hard attention.

In the up-sampling path of the U-Net, the spatial information recreated from the activation feature maps is incorrect. To counteract this, U-Net uses skip connections joining spatial information from the down-sampling path with the corresponding activation feature maps of the up-sampling path. This is because feature maps from the down-sampling path have very good spatial information but very poor feature quality, whereas feature maps from the up-sampling path have very good feature representation as they come from a deeper part of the network but have poor spatial information. Thus, implementing skip-connections gives the best combination of spatial and feature representation.

However, implementing the skip connection brings over redundant features from the down-sampling blocks because they have poor feature quality as they are

extracted from shallower layers of the network. Placing AGs at the spot of the skip connection helps us suppress this redundant information. In Figure 2.2, we can see that there are gates placed in places where the skip connections are implemented.

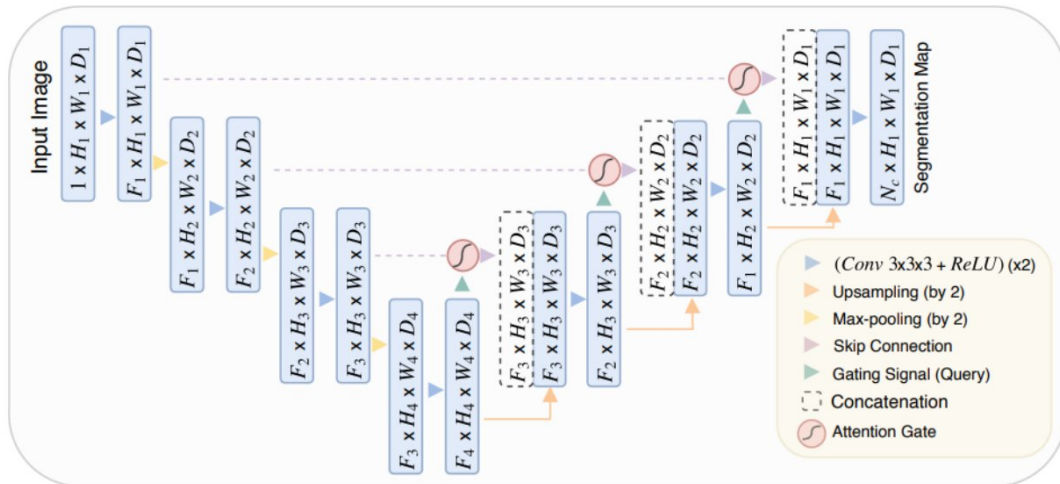


Figure 2.2: Attention U-Net Architecture (Courtesy of [39])

The authors of [39] show that incorporating AGs in the skip connections improves the performance across all the metrics. The results of Attention U-Net compared to U-Net on the TCIA Pancreas-CT Dataset [14] is shown in Table 2.1.

	Method	Dice Score	Precision	Recall	S2S Dist (mm)
<u>BFT</u>	U-Net	0.690 ± 0.132	0.680±0.109	0.733±0.190	6.389±3.900
	Attention U-Net	0.712±0.110	0.693±0.115	0.751±0.149	5.251±2.551
<u>AFT</u>	U-Net	0.820±0.043	0.824±0.070	0.828±0.064	2.464±0.529
	Attention U-Net	0.831±0.038	0.825±0.073	0.840±0.053	2.305±0.568
<u>SCR</u>	U-Net	0.815±0.068	0.815±0.105	0.826±0.062	2.576±1.180
	Attention U-Net	0.821±0.057	0.815±0.093	0.835±0.057	2.333±0.856

Table 2.1: Attention U-Net vs. U-Net on the TCIA Pancreas-CT Dataset in all three conditions: Before Fine Tuning (BFT), After fine tuning(AFT), and Models trained from scratch (SCR) (Courtesy of [39])

2.3 Double U-Net

Debesh Jha et. al. [29] proposed Double U-Net, a novel architecture for semantic image segmentation. Although Attention U-Net improved upon the traditional U-Net, it still remained a simple convolutional architecture. To capture more complex regions of interest, the authors proposed an architecture that uses two U-Nets in sequence. The Double U-Net outperforms the standalone U-Net due to a lot of factors. A typical architecture of the Double U-Net is shown in Figure 2.3. The improvements of Double U-Net over the standalone U-Net are achieved with the help of the following modules:

- **Encoder 1:** Instead of a normal encoder, the authors use a VGG-19 architecture [42] pre-trained on the Imagenet [16] dataset. The intuition is that its architecture is quite similar to that of the encoder path of a U-Net. On top of that, it harnesses transfer learning to leverage the pre-trained weights while training on small datasets.
- **Atrous Spatial Pyramid Pooling (ASPP) Layer:** The ASPP [11] block uses re-sampling techniques at multiple rates and dilated convolutions to extract more meaningful contextual information compared to other context extraction techniques.

The ASPP block is used in both the bottleneck layers of Network 1 and Network 2.

- **Concatenation of the output from Network 1 and Network 2:** If we concatenate the outputs from network 1 and network 2 the boundaries of the segmentation mask get much more refined, which is visible in Figure 2.4.
- **Squeeze and Excite blocks:** As the network gets deeper, the number of channels increases. Each of these channels might contain different types of feature representation. Thus, each channel of the model's filters can learn different weights. The task of the Squeeze and Excite blocks[25] is to filter

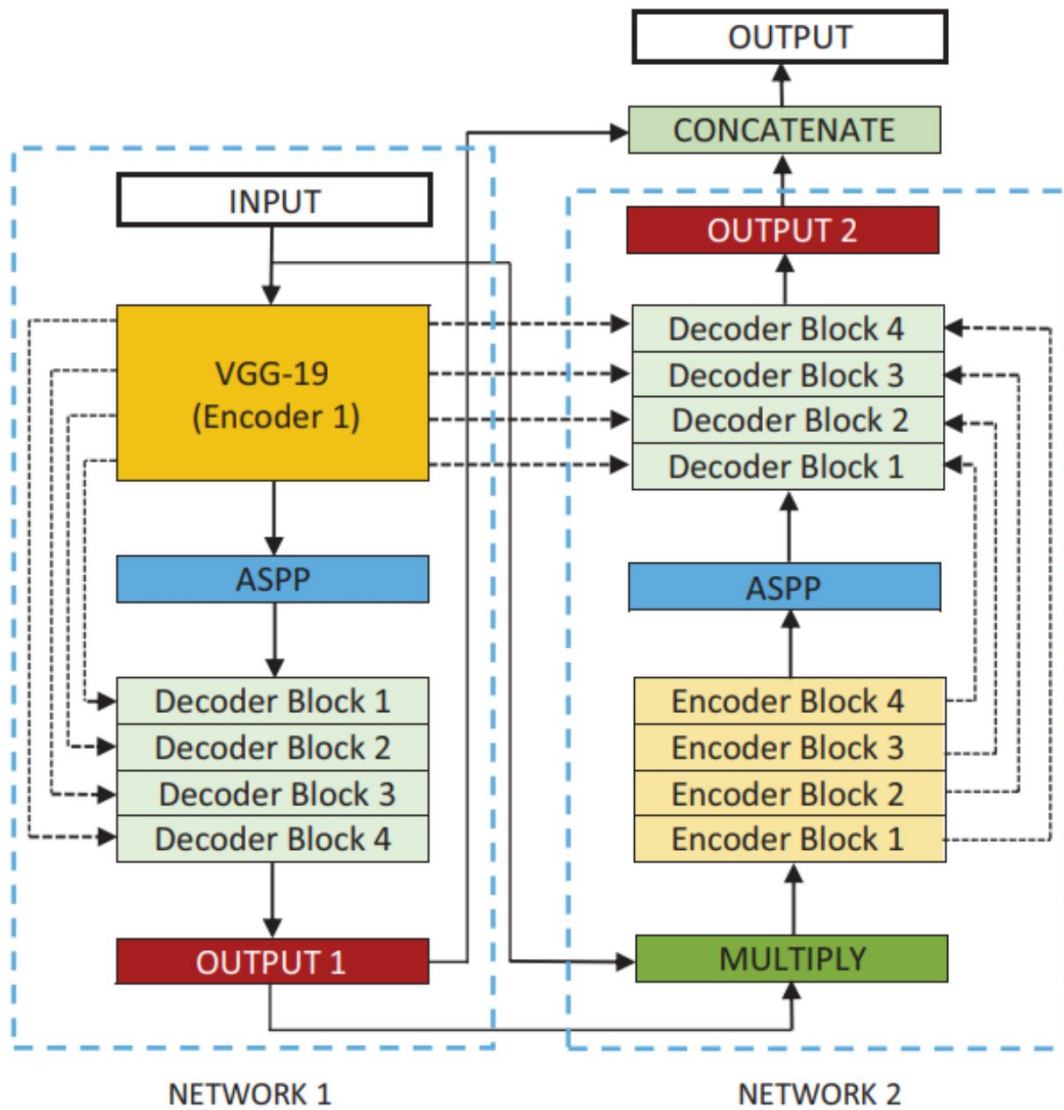


Figure 2.3: Block diagram of the Double U-Net Architecture (Courtesy of [29])

out which channel information is more relevant by using a series of fully connected layers followed by ReLU activation. There are Squeeze and Excite blocks in each of the blocks of decoder 1, encoder 2, and decoder 2 of the Double U-Net.

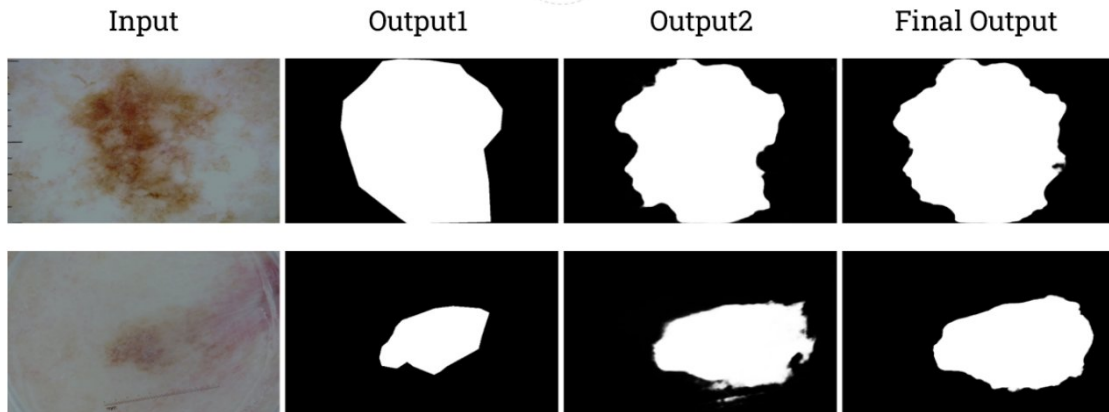


Figure 2.4: Impact of concatenating two outputs of the Double U-Net (Courtesy of [29])

Double U-Net performed significantly better than other baseline models. The authors performed experiments on the CVC-Clinic DB dataset [4] for polyp segmentation, the ISIC 2018 Dataset [15] for skin lesion segmentation and the 2018 Data Science Bowl dataset [7] for nuclei segmentation. They achieved a DSC score of **0.9239** on the polyp dataset surpassing the traditional U-Net which scored 0.8781. For the lesion dataset, they outperformed other state-of-the-art methods by **5.7%**. Finally, for the nuclei segmentation, the authors achieved a promising DSC score of **0.9133**. These results show that the modules discussed above effectively improve Double U-Net over its traditional counterpart.

2.4 Skin Lesion Segmentation using Generative Adversarial Networks

Izadi et. al. [27] proposed a pix2pix [26] based segmentor network to segment out skin lesions. They used the U-Net as a generator and used a Critic network

which acts as the discriminator, by assigning a real value number to the input of the discriminator. The architecture of the network proposed in this paper is shown in Figure 2.5. The authors trained the model on the DermoFit dataset [2] and achieved a DSC score of **0.898** improving on the traditional U-Net’s score of 0.887, thus showing that adding a critic network along with U-Net produces better segmentation masks.

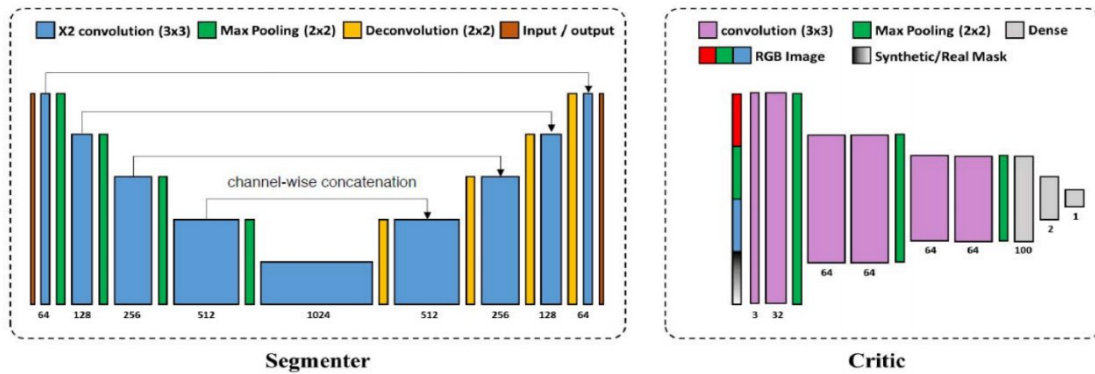


Figure 2.5: Architecture proposed by Izadi et. al. (Courtesy of [27])

Wei et. al. [48] proposed a GAN-based boundary aware architecture as shown in Figure 2.6 where they tackled the challenge of irregular boundaries by introducing a Scale-Att-ASPP module in the skip connections for more contextual information. They also introduced a multi-scale L1 loss that guides the model to learn more meaningful boundary information. Their proposed model was evaluated on the ISIC 2017 dataset [5] and they performed slightly better compared to other state-of-the-art models at the time with a dice coefficient of **0.8781**.

2.5 Transformer-based model for medical image segmentation

Convolution operations are generally good for modeling local information, hence U-net [41] works very poorly when trying to model long-range dependencies. As a result, U-Net is not able to segment out structures that are not regular and shows a great range of variation. Chen et. al. [10] proposes a new architecture called the

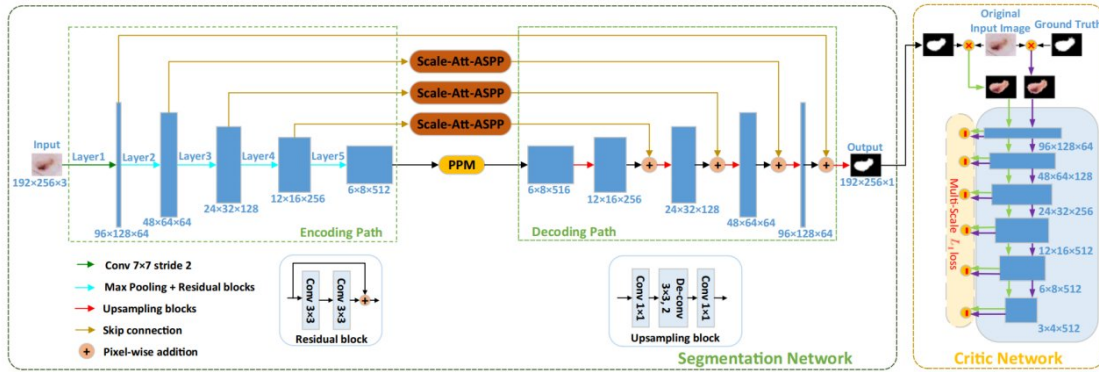


Figure 2.6: Architecture proposed by Wei et. al. (Courtesy of [48])

TransUNet which uses both U-Net and Transformers that solves the problem of U-Net alone not being able to capture long-range dependencies. In their proposed architecture the Transformers act as a strong encoder.

If we directly use the transformer as an encoder on the tokenized input image and use decoders to upsample the encoded features, it does not produce satisfactory segmentation maps. As Transformer treats its input as 1D sequences and thus concentrates on how to model global context at all stages, this results in low-resolution features lacking detailed localization information. To alleviate this problem the authors of this paper propose a slightly different approach to designing the encoding path by using a hybrid CNN-Transformer architecture. This leverages both the high-resolution features provided by the CNN and the global context extracted by the Transformer. In 2.7, initially in the encoding path, CNN is used to extract high-resolution features from the input image. After that 1x1 patches are extracted from the feature maps of the CNN, and patch embedding is applied to the extracted patches, which maps the vectorized patches into a D-dimensional embedding space using a trainable linear projection. The decoding path consists of multiple upsampling layers taking in encoded features and through a series of convolution and transposed convolution the required segmentation map is generated.

For experimentation, the authors trained their model on the Synapse Multi-Organ

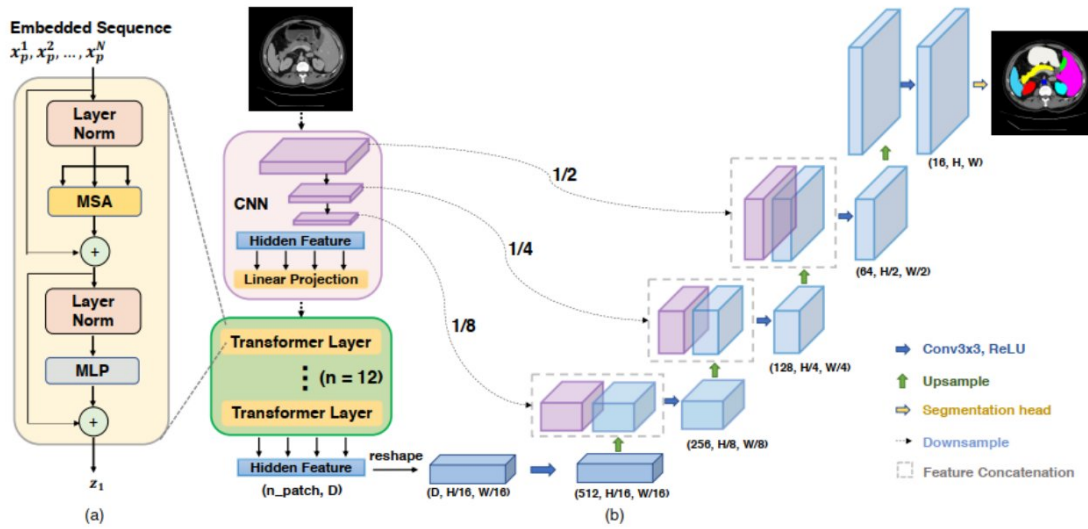


Figure 2.7: Architecture of TransUnet. (a) block diagram of the transformer layer, (b) overall framework of TransUnet (Courtesy of [10])

segmentation dataset ¹ and the Automatic cardiac diagnosis challenge (ACDC) dataset ². They achieved an average Dice Score of **0.7748** on the Multi-Organ dataset and **0.8971** on the ACDC dataset, improving well upon the likes of U-Net, Attention U-Net and ViT itself.

2.6 Polar image transformations to improve medical image segmentation

Rather than proposing an improved model for image segmentation tasks, Marin Bencevic et. al. [3] proposed a pre-processing technique that improves neural network performance and data efficiency on segmentation tasks. Their suggested method was to convert cartesian images to polar coordinates such that the center point of the object becomes the polar origin for the transformation. This reduces the dimensionality of the image and also separates the segmentation task from the localization task which helps the model to converge easier. They proposed two methods for obtaining the polar origin of an image: (1) Estimating with a model

¹<https://www.synapse.org/#!Synapse:syn3193805/wiki/217789>

²<https://www.creatis.insa-lyon.fr/Challenge/acdc/>

trained on non-polar images, and (2) Estimating with a model trained to predict the optimal origin.

In the first method, the authors use a cartesian network that takes in a cartesian image as input and performs an initial segmentation. This cartesian network can be any neural network of choice that can perform segmentation tasks. The authors of this paper used U-Net [41], Res-U-Net++ [31], and DeepLabV3+ [12] as their choice of neural networks for all the experiments. Using the segmentation produced by the cartesian network, the polar origin or center of mass of the image can be calculated using a few simple mathematical equations. To calculate the center of mass of an image $I(x, y)$, at first the spatial image moments matrix M needs to be determined. The entry of the matrix at the i -th row and j -th column can be calculated using the following formula:

$$M_{i,j} = \sum_{x,y} I(x,y).x^i.y^j \quad (2.1)$$

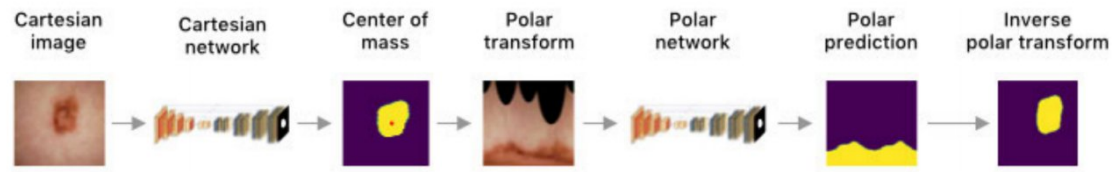
After this, the center of mass of the image (c_x, c_y) can be calculated using the following formula:

$$c_x = M_{10}/M_{00} \quad (2.2)$$

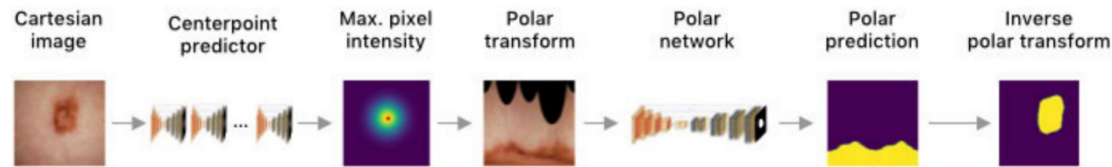
$$c_y = M_{01}/M_{00} \quad (2.3)$$

This center of mass is used to transform the original cartesian input image to its corresponding polar coordinates. This polar image is then fed into the polar network which is another neural network of choice. This network produces a second segmentation mask in polar coordinates and after applying the inverse polar transformation to this output, the final cartesian segmentation mask is obtained. A walk-through of this method can be seen in Figure 2.8a.

In the second method, a center-point predictor network is trained whose task is to specifically predict the center of mass of the input image. This model is based on the stacked hourglass architecture [37]. The model takes a cartesian image



(a) Proposed pipeline that uses a neural network to predict center of mass and then another polar network to produce the polar prediction



(b) Proposed pipeline that uses an hourglass architecture as center-point predictor and then a polar network to produce the polar prediction

Figure 2.8: Two of the proposed methodologies illustrating the pipelines for image segmentation using polar transformation (Courtesy of [3])

as input and generates a heatmap prediction of the image. The ground truth heatmaps required for training the model are generated using the same equations discussed in 2.2. The coordinate of the pixel with the highest intensity in the heatmap is considered to be the polar origin which is then used to transform the cartesian image to its polar form. The polar image is then fed to the polar network similar to the previous method. The polar network produces a polar segmentation mask on which inverse polar transformation is applied to generate the final output segmentation mask. The steps of this method can be seen in Figure 2.8b.

The authors of [3] show in their experiments on the polyp [4] and skin lesion [15] datasets that they got the best results using the second method of the center-point predictor. They were able to achieve a dice coefficient of **0.9374** on the polyp dataset using U-Net as their choice of polar network and a dice coefficient of **0.9253** on the skin lesion dataset using Res-U-Net++ as the polar network making their approach the current state-of-the-art.

CHAPTER 3

PROPOSED METHODOLOGY

The traditional U-Net has a very simple structure of a decoding path followed by an encoding path. In the decoding path, the activation cannot be up-sampled properly as spatial information gets lost. To overcome this, skip connections are added from the encoder blocks to the decoder blocks. But due to poor representation of features from the encoding path, irrelevant features are also concatenated in the skip connections. The Attention U-Net alleviates this problem by discarding irrelevant features using attention gates. The Double U-Net proposes using a pre-trained VGG-19 model as the first encoder, ASPP layers, Squeeze and Excite blocks, and finally concatenating the outputs of two U-Nets stacked on top of each other to produce more refined segmentation maps. However, it still fails to filter out the irrelevant features being concatenated in the skip connections. Thus, the Double U-Net leaves scope for spatial attention to be implemented and filters out those irrelevant features through the skip connections.

The Double U-Net is a very large network with a significant number of parameters and the architecture proposed by the authors does not include residual connections in the encoding and decoding blocks. To ensure better gradient flow and quick convergence, we consider incorporating residual connections in the encoding and decoding blocks.

Finally, the authors of Double U-Net suggest augmenting the ISIC 2018 dataset to increase the dataset size to 50,000 images. Unfortunately, this is computationally infeasible for us to train. Thus, it provides us an avenue to use the color constancy algorithm as a pre-processing step that can reduce the number of training images in the dataset making the training process much more computationally feasible.

With these research gaps in mind, we propose a spatial attention-based Double U-Net architecture with residual connections across the convolutional blocks that can be trained using a smaller dataset using a color constancy algorithm as a pre-processing technique.

3.1 Overview of our proposed architecture

Figure 3.1 illustrates the block diagram for our proposed architecture. The input image is fed into Encoder 1 which is a pre-trained VGG-19 network, followed by an ASPP block which is used to retain contextual information. The activation maps produced are then passed to Decoder 1 which is a series of upsampling convolutional blocks to generate an output. This result is further multiplied with the input image and then passed through another set of Encoder-ASPP-Decoder path to produce a second output. The outputs from the first and second decoder are concatenated and passed through a final convolutional block to generate our final segmentation mask as the output of our model. We incorporate attention-based skip connections between Encoder 1-Decoder 1, and Encoder 2-Decoder 2 to retain more spatial information.

As shown in Figure 3.7b the convolutional blocks that make up each of the encoders and decoders contain residual connections between the input of the convolutional block and the output of the CONV-BATCHNORM-ReLU-CONV-BATCHNORM path to better learn the identity function to prevent vanishing gradients. Furthermore, dropout is added as means of regularization to prevent overfitting. Finally, the output is passed through a squeeze and excite block that provides channel-wise attention and the final output is fed to a subsequent convolutional block.

We apply color constancy as a pre-processing technique to the input images before feeding them to our network. Alongside this, our proposed model has four main components, namely a squeeze and excite block, an ASPP block, attention gates on skip connections and a modified convolutional block with residual connections. The data pre-processing and key components are discussed in detail for the rest of this chapter.

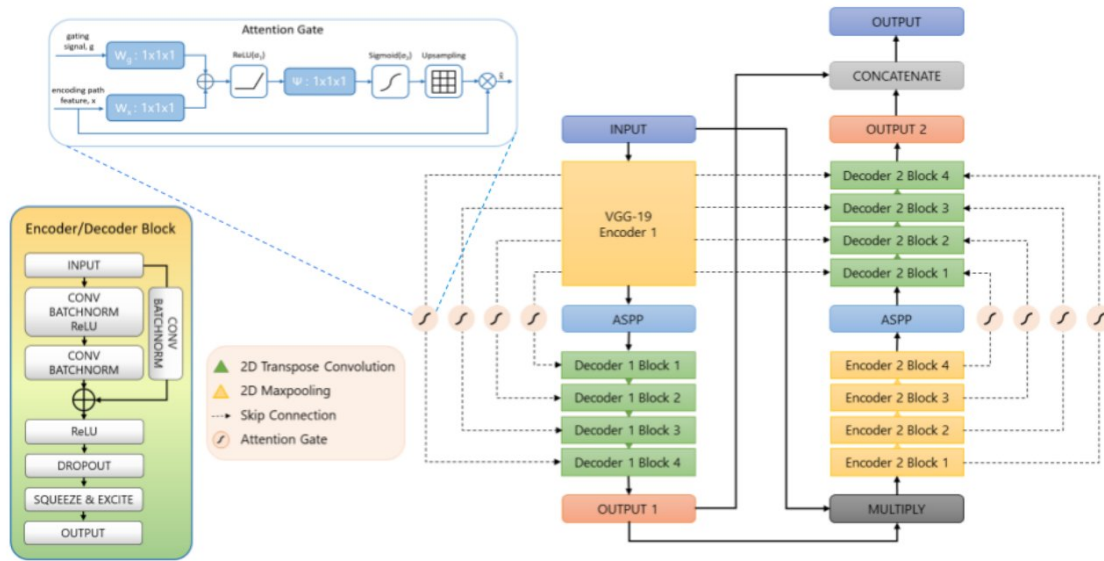


Figure 3.1: Block diagram of the proposed model architecture

3.2 Data preprocessing: Color Constancy

Dermoscopy and colonoscopy images are captured on different camera devices and consequently under different light sources. There are also places of high reflectance in the images containing polyp. All these issues can be alleviated if we can normalize the image using the Color Constancy (CC) [38] as a pre-processing step. Even though there are changes in the illumination of different pictures, color constancy helps bring the illumination of all the images into the same illumination spectrum, which in most cases is the white light.

In Figure 3.2, we observe that before applying color constancy the values in the

red channel are higher than all the other channels. As a result the image also looks reddish. After the pre-processing is done we see all the channels have an equal value and produce an image with near-white skin as seen on the bottom left.

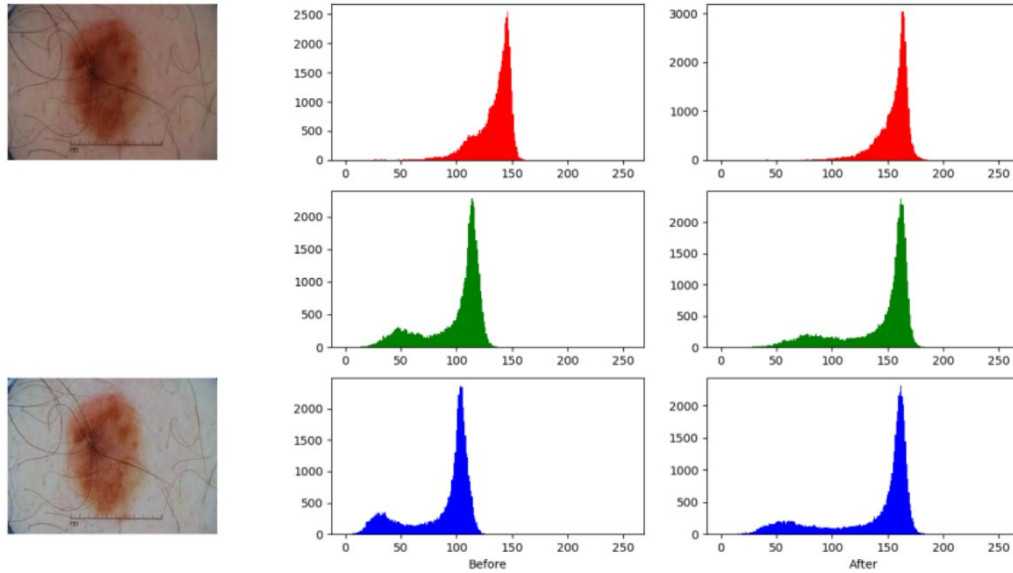


Figure 3.2: The graphs in the middle column show the values of each RGB color channel before normalization, where it can be clearly seen that R channel values are higher, resulting in the red skin of the original image shown in top left. After normalizing each channel value, the values for each channel get distributed to similar values, producing an image with near white skin as seen on the lower left. (Courtesy of [38])

There are several algorithms for color constancy but the one that we used is the shades of gray algorithm [20]. This algorithm is formed from the notion that color constancy would perform better if the scene average, that is, the color to which different images are brought is a shade of grey. The equation of the algorithm is shown below:

$$\left(\frac{\int f(x)^p dx}{\int dx} \right) = K_e \quad (3.1)$$

Here p is the Minkowski Norm and shades of grey works best with $p = 6$. An example of shades of grey is shown below in Figure 3.3.

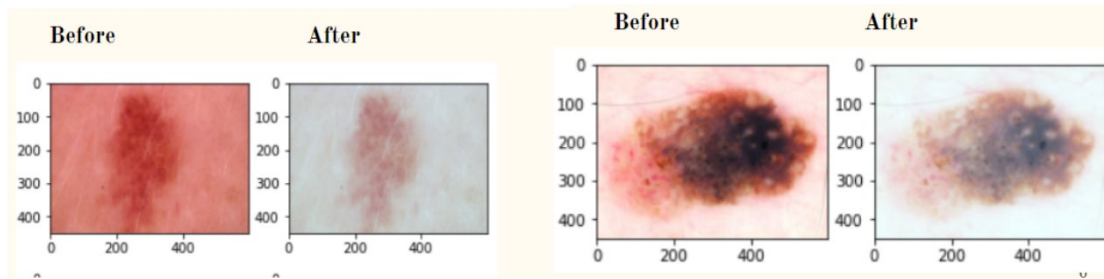


Figure 3.3: Before and after applying shades of gray variant

For the ISIC 2017 dataset [5], the authors Hua Ng et. el. [38] found that applying the shades of gray [20] algorithm improved the Dice Coefficient by at least 1 percent for three classes Nevi, Melanoma, and Seborrheic Keratosis. Moreover, this is one of the easiest color constancy algorithms to implement. Thus, we got the motivation to use the shades of gray algorithm as a pre-processing step in our methods.

3.3 Squeeze and Excite Block

Double U-Net employs the Squeeze and Excite blocks which recalculates the channel-wise feature responses by modeling the dependencies between the activation in each of the channels. In Figure 3.4, we can see that the height and width of the tensor are made to be one which is then passed through a series of fully connected blocks and activation. As the input propagates through this network, it filters out which channel information is more relevant. By using this Squeeze and Excite module in each of the blocks of decoder 1, encoder 2, and decoder 2, we can construct an architecture that generalises extremely well across challenging datasets.

3.4 Atrous Spatial Pyramid Pooling (ASPP)

ASPP is a module that performs re-sampling on activation maps at different rates before applying convolutions. We apply multiply filters that have complementary effective receptive fields, capturing relevant contextual information from the input

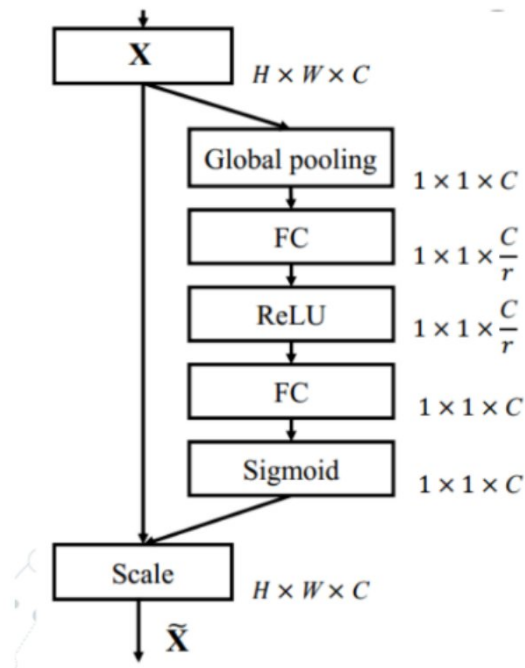


Figure 3.4: Squeeze and Excite Block Architecture (Courtesy of [25])

image at multiple scales. This re-sampling is implemented using multiple parallel dilated convolutional layers with different sampling rates [50]. Figure 3.5 gives an illustration of how ASPP exploits multi-scale features by applying multiple parallel filters to classify the center orange pixel.

3.5 Attention Gates

As channel-wise attention is already employed in Double U-Net, there is room for improvement by incorporating spatial attention to suppress the irrelevant features coming from the down-sampling path in the skip connections. Attention gates help the model focus on relevant activations by giving prioritizing relevant regions more than non-relevant regions. AGs do this by element-wise multiplying the input coming from the encoder path with a weight matrix generated by the attention gate. The detailed diagram of the gate is shown in Figure 3.6.

On the basis of Figure 3.6, the following describes how AGs work:

- The gate takes in two inputs \mathbf{x} and \mathbf{g} , which have varying size as \mathbf{x} comes

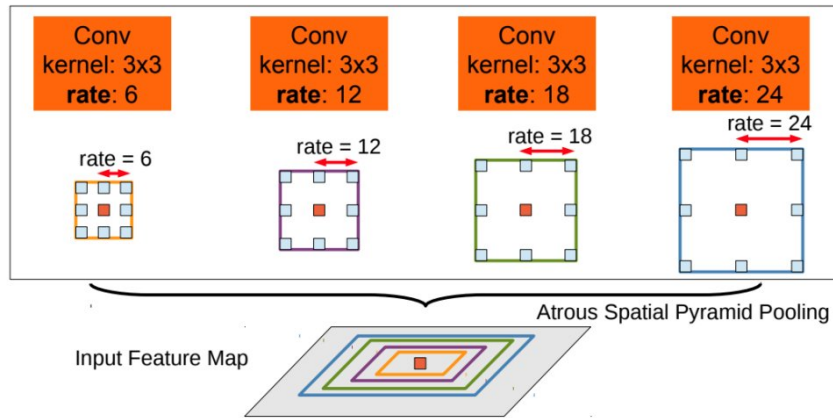


Figure 3.5: ASPP employs dilated convolutions with different rates to exploit multi-scale features and classify the orange pixel (Courtesy of [11])

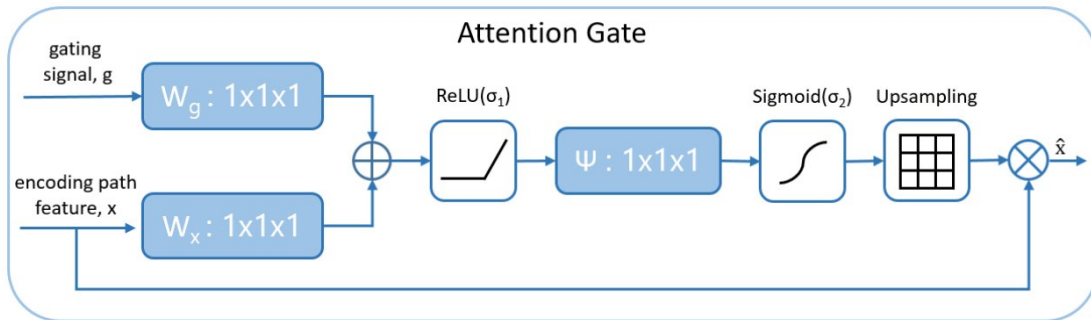


Figure 3.6: Attention Gate structure (Courtesy of [39])

from the down-sampling path one level higher compared to the gating signal \mathbf{g} .

- Let's consider \mathbf{x} to have a size of $[128 \times 128 \times 128]$ and \mathbf{g} to have a size of $[64 \times 64 \times 64]$.
- To make the size of the two signals equal, \mathbf{x} goes through a convolution with filter size (1×1) , stride=2 and number of filters=128. \mathbf{g} goes through a convolution with filter size (1×1) , stride=1 and number of filters = 128. Thus both \mathbf{x} and \mathbf{g} have the same size of $[64 \times 64 \times 128]$.
- The two signals of equal size are then added which makes the aligned weights larger and unaligned weights comparatively smaller.

- The resultant signal passes through a 1x1 convolution with number of filters=1, which collapses the size to [64x64x1].
- This signal then passes through a sigmoid activation which brings the activation values between 0 $\tilde{1}$. This produces weights to filter out irrelevant information coming from the down-sampling path. A weight closer to 1 indicates that particular activation is important and vice versa.
- The resultant signal is up-sampled to the size of \mathbf{x} and element-wise multiplication is done between \mathbf{x} and the activation weights. Thus we are multiplying each pixel of \mathbf{x} with a weight calculated from the Attention Gate. These weights get updated after each epoch through back-propagation.
- Through the multiplication, the dominant features are retained as they are multiplied with values close to 1, while the irrelevant features are neglected as they get multiplied with values close to 0.

Our intuition behind incorporating AGs is that since the traditional Attention U-Net generates better feature maps with the help of AGs in skip connections, the Double U-Net could leverage this soft attention mechanism as well. Hence we incorporated AGs in the places where skip connections occur in the Double U-Net.

3.6 Residual Connections

Double U-Net is a very large network as it is one U-Net, which by itself is a neural network, stacked on top of another. Furthermore, there are additional components like the Squeeze and Excite blocks and ASPP blocks. So quite naturally, the network gets very deep with a substantial number of parameters. As deeper networks fail to propagate small changes in derivatives to earlier layers, they suffer from vanishing gradient problem. To help facilitate proper gradient flow and smoother convergence of the network, a need for residual connections arises. As explained in [24], residual connections help deep networks perform better and converge faster. Our intuition was to employ residual connections in the encoder and

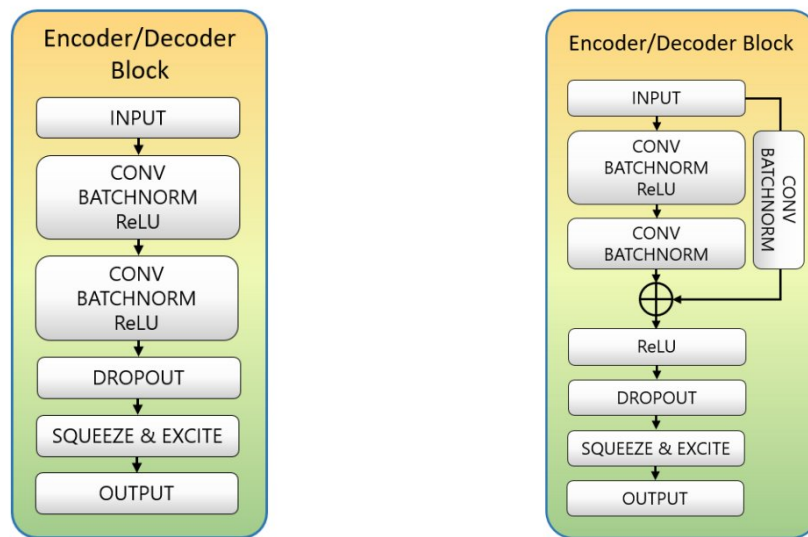
decoder block units as Double U-Net is a deep architecture.

3.7 Improvement of our proposed convolutional block over that of Double U-Net

In Figure 3.7, we can observe the differences between the structure of the convolution blocks of Double U-Net and our proposed model. The convolution blocks contain the fundamental components that build up each individual block of decoder 1, encoder 2, and decoder 2. Figure 3.7a shows the components of the convolution block of the original Double U-Net. It consists of two consecutive CONV-BATCHNORM-ReLU operations followed by a Dropout layer that proceeds to a Squeeze and Excite block. Figure 3.7b illustrates the changes that we made to the convolution block in our proposed model. Here, the input passes through two separate paths. The first path is similar to the original Double U-Net block where the input passes through a series of CONV-BATCHNORM-ReLU-CONV-BATCHNORM operations sequentially to produce an intermediate output. But there is another parallel path where the same input goes through a single CONV-BATCHNORM operation before being added to the intermediate output produced from the first path. The addition of these two paths forms the residual or skip connection that improves the gradient flow in our proposed model. The concatenated output is then passed through a ReLU activation which is followed by a Dropout layer before passing through a Squeeze and Excite block.

3.8 Architecture Variants

So far, we described the addition of residual connections and AGs in the Double U-Net architecture and the entire structure of the model that we propose is shown in Figure 3.1. We added AGs in the skip-connections between encoder 1-decoder 1, as well as between encoder 2-decoder 2. Based on the placement of attention gates, we propose two variants of our model: Half-Attention Double U-Net and



(a) Double U-Net Convolution Block (b) Our Proposed Convolution Block

Figure 3.7: Comparison of the convolutional blocks of the Double U-Net with ours

Full-Attention Double U-Net. When attention gates are placed only in the skip connections between encoder 1-decoder1, we call the model Half-Attention Double U-Net as shown in Figure 3.8.

When AGs are placed in the skip-connections between encoder 1-decoder 1 as well as the skip-connections between encoder 2-decoder 2, we call it the Full-Attention Double U-Net as shown in Figure 3.9.

3.9 Loss Function

To train and evaluate our model on each of the datasets we chose Dice loss as our loss function which is defined by the following formula.

$$Dice_{loss} = 1 - \frac{2|Image_{pred} \cap Image_{gt}| + \lambda}{|Image_{pred}| + |Image_{gt}| + \lambda} \quad (3.2)$$

Here, λ represents a small constant to prevent dividing-by-zero. $Image_{pred}$ and $Image_{gt}$ refers to the predicted mask and ground truth mask respectively.

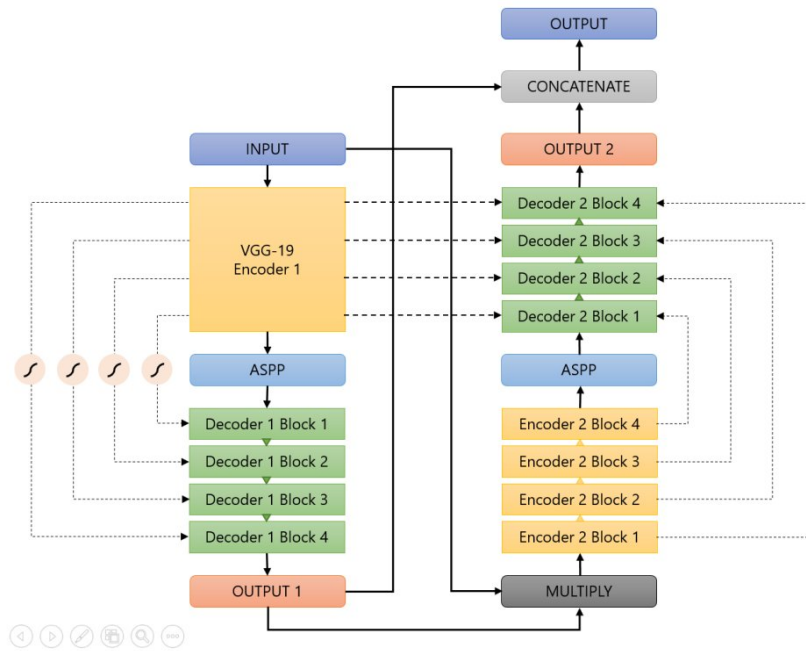


Figure 3.8: Block diagram of the Half-Attention Double U-Net

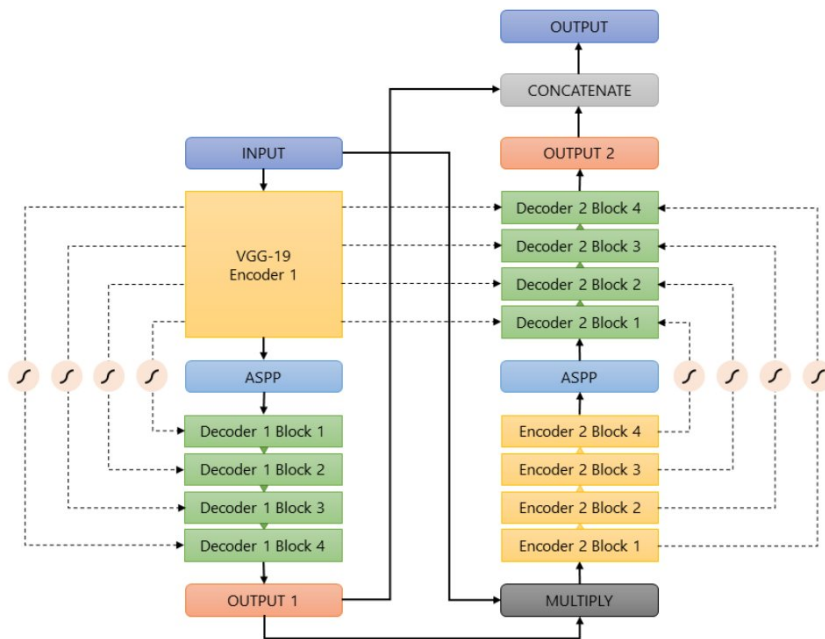


Figure 3.9: Block diagram of the Full-Attention Double U-Net

CHAPTER 4

RESULT ANALYSIS

4.1 Dataset Description

There are several datasets in the medical image domain that are regarded as benchmark datasets. Among these datasets, we used the ISIC2018 (Skin Lesion), CVC-ClinicDB (Polyp), and 2018 Data Science Bowl (Nuclei) datasets for training, validating, and testing our proposed methods and architectures.

4.1.1 ISIC 2018

The ISIC 2018 Challenge [15] on Skin Lesion Analysis Towards Melanoma Detection was divided into three parts among which the first task was skin lesion segmentation. This dataset includes 2594 dermoscopic images along with corresponding professional annotated binary segmentation masks. Some samples of this dataset can be seen in Figure 4.1.

4.1.2 CVC-ClinicDB

CVC-ClinicDB [4] is a database containing colonoscopy images and their corresponding binary segmentation masks. These masks correspond to the region of

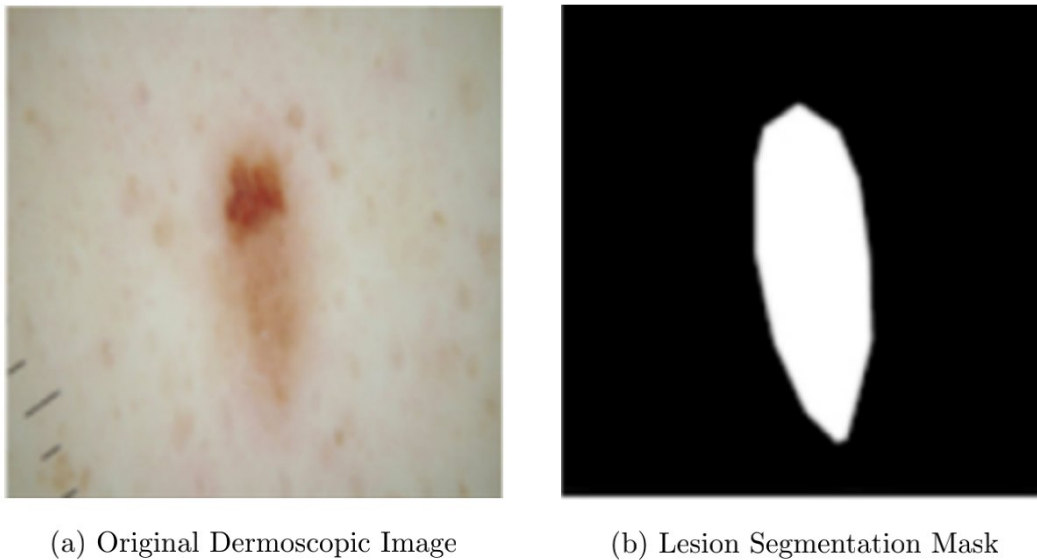


Figure 4.1: Sample Image and Corresponding Binary Segmentation Ground Truth Mask from ISIC 2018 Dataset (Courtesy of [15])

the image covered by polyps. The colonoscopy images were extracted as frames from colonoscopy video sequences. Several types of polyps can be found within these frames. This dataset contains 612 images from 29 video sequences along with the manually annotated ground truth mask covering the polyp associated with each image. The images have a resolution of 388x288 pixels. We can see a few examples of these colonoscopy images along with their corresponding ground truths in Figure 4.2.

4.1.3 2018 Data Science Bowl

The 2018 Data Science Bowl [7] dataset contains segmented nuclei images acquired under a variety of conditions such as cell type, magnification, and imaging modality (brightfield vs. fluorescence). This dataset forms a diverse collection of biological images containing tens of thousands of nuclei. The nuclei in the images are derived from a range of organisms including humans, mice, and flies. The nuclei also appear in different contexts and states including cultured mono-layers, tissues, embryos, cell division, genotoxic stress, and differentiation. This dataset is designed in such a way that it challenges a model's generalizability across these

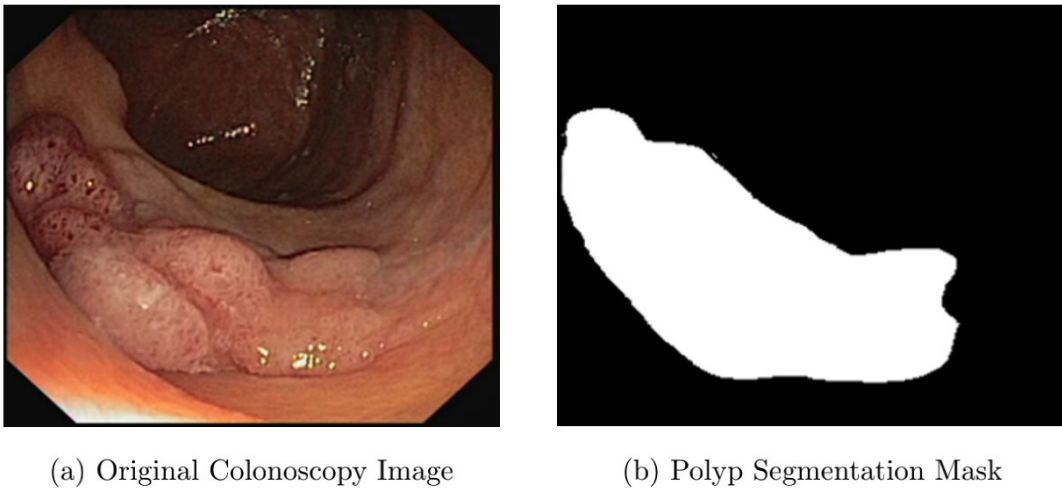


Figure 4.2: Example of Colonoscopy Image from CVC-ClinicDB Database Along with its corresponding Ground Truth (Courtesy of [4])

variations. It contains 670 nuclei images and the segmented masks of each nucleus. Each mask contains one nucleus and is not allowed to overlap, i.e. no pixel belongs to two different masks. A few example images of this dataset can be seen in Figure 4.3.

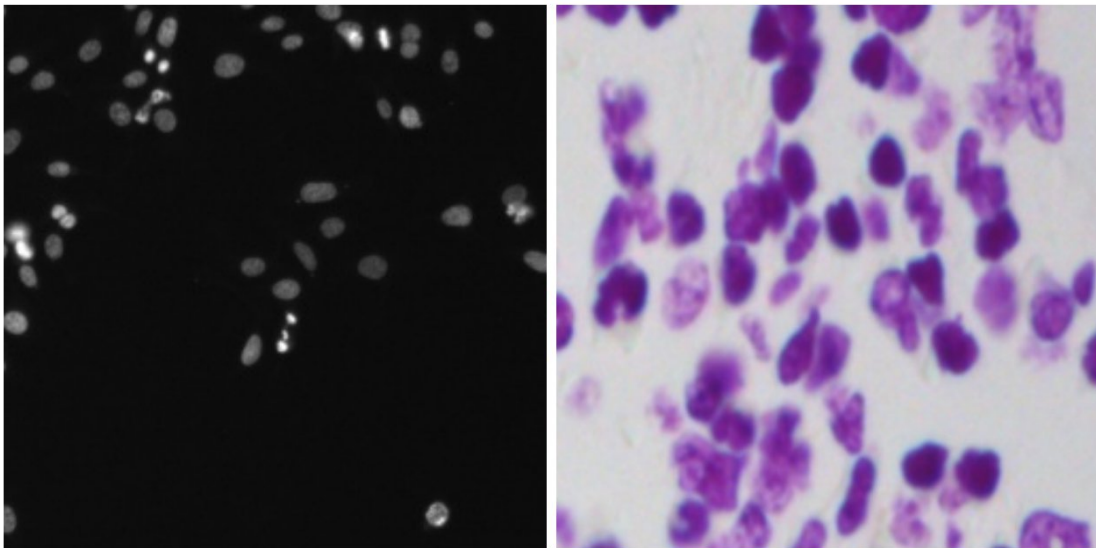


Figure 4.3: Example Nuclei Images from 2018 Data Science Bowl Dataset (Courtesy of [7])

4.2 Experimental Setup

We trained our model and performed all experiments on Kaggle using Tesla V100 GPU. The training hyper-parameters used in our experiments for each dataset are shown in Table 4.1. All of the datasets were split into an 80-10-10 train-valid-test split. For the polyp and nuclei dataset, we used the original image size since they are small datasets. However, for the lesion dataset, we had to resize the images to 192*256 to balance between training time and complexity. We used Dice Loss as the loss function and the Nadam [18] optimizer for each dataset. The learning rates were varied according to the size of the dataset. We used a lower learning rate for the nuclei dataset since the model would run out of data before convergence.

Dataset	ISIC 2018	CVC Clinic	Nuclei
Image Size	192*256	288*384	256*256
Loss	Dice Loss	Dice Loss	Dice Loss
Optimizer	Nadam	Nadam	Nadam
Learning Rate	0.0001	0.0001	0.00001
Epochs	40	35	30

Table 4.1: Training hyperparamaters for each dataset

Alongside applying CC, we performed normalization and sample wise centering on the images to ensure zero mean and unit standard deviation. Random rotation, Vertical and horizontal flip, converting to Hue Saturation Value (HSV) form, random brightness-contrast, and histogram equalization were among the data augmentation techniques we applied to each of the datasets to increase the dataset sizes by six-fold [6].

4.3 Evaluation Metrics

The measure of dice coefficient is the standard evaluation metric for skin lesion segmentation. It represents the similarity between the generated skin lesion regions and those of the ground truth. Apart from this, we have used other evaluation metrics including Intersection over Union (IOU), precision, and recall which are associated with four values, i.e. true-positive(TP), true-negative(TN), false-positive(FP), and false-negative(FN). These four metrics are widely used in image segmentation literature, and therefore are suitable choices for our experiments.

$$Dice = \frac{2 * TP}{(TP + FP) + (TP + FN)} \quad (4.1)$$

$$IOU = \frac{TP}{TP + FP + FN} \quad (4.2)$$

$$Precision = \frac{TP}{TP + FP} \quad (4.3)$$

$$Recall = \frac{TP}{TP + FN} \quad (4.4)$$

4.4 Quantitative Results

4.4.1 Ablation study

We conducted three experiments incorporating different modules to determine which overall pipeline works best for our cause. The first experiment is concerned with the comparison between the two variants of our architecture. One is the Half Attention Double U-Net, where the attention gates are only placed between the first encoder and the first decoder, whereas the Full Attention variant also includes attention gates between the second encoder and decoder. Table 4.2 indicates that the Full Attention variant results in better performance (DSC of **91.64%**) compared to its counterpart (DSC of **90.9%**). This implies that adding more attention gates in the skip connections allows for more relevant spatial information to be harnessed providing better segmentation results.

Model	DSC(%)	IOU(%)	Precision(%)	Recall(%)
Half-Attention-DU-Net	90.9	83.38	94.63	85.79
Full-Attention-DU-Net	91.64	84.63	95.76	86.13

Table 4.2: Result of our proposed approaches on ISIC 2018 dataset

The goal of our second experiment was to determine the effectiveness of adding a critic network in improving the results of our proposed architecture. The job of a critic network in a conditional GAN is to look at each individual patch in the image generated by the generator and evaluate how real or fake they are. Using this feedback from the critic, the generator can adjust its weight overtime to generate better and better segmentation masks. As we can see from Table 4.3, initially, we found that training the baseline DU-Net as the generator in a conditional GAN setting helps to improve the DSC from **89.62%** to **89.70%**. However, when we tried to use our Full-Attention-DU-Net as the generator, the model started to overfit (training DSC **97.68%** and test DSC **84.10%**) due to a large number of added parameters and a small amount of training data. Consequently, we removed the critic network and trained the Full-Attention-DU-Net model separately which yielded a DSC of **91.64%**.

Model	DSC(%)
DU-Net	89.62
DU-Net + Critic Network (cGAN)	89.70
Full-Attention-DU-Net + Critic Network (cGAN)	84.10
Full-Attention-DU-Net	91.64

Table 4.3: Comparative result of our proposed model with and without a critic network ISIC 2018 dataset

Our aim in the third experiment was to illustrate the effectiveness of the different modules that we wish to incorporate. One is the pre-processing of images using color constancy, and the other is the incorporation of residual connections in the convolution blocks of the encoder/decoder networks. As shown in Table 4.4, we

observe that applying CC improves our model from a DSC of **90.73%** to **91.64%**. This shows that transforming images so that they appear under a uniform light source helps the model to achieve better segmentation results. Moreover, adding residual connections further enhance the DSC to **91.68%** which leads us to the conclusion that the Full Attention Double U-Net model with residual connections across the convolution blocks trained on images pre-processed by applying CC generates the best possible results for the ISIC 2018 benchmark dataset.

Model	DSC(%)	IOU(%)	Precision(%)	Recall(%)
Full-Attention-DU-Net (Without CC)	90.73	83.14	92.54	88.66
Full-Attention-DU-Net (CC)	91.64	84.63	95.76	86.13
Full-Attention-Res-DU-Net (CC)	91.68	84.68	94.19	87.55

Table 4.4: Results of our proposed model on the effect of adding color constancy and residual connections on ISIC 2018 dataset

4.4.2 Training progression

We illustrate the learning process of our model through the graphs presented in Figure 4.4, 4.5 and 4.6 where we show how the DSC scores and Dice losses converge when training on all three datasets. In each of the cases, we can observe the training and validation DSC scores are close to each other at the end of training and have plateaued off except the one for the polyp dataset in Figure 4.5 where we can observe a difference in train and valid loss indicating slight overfitting.

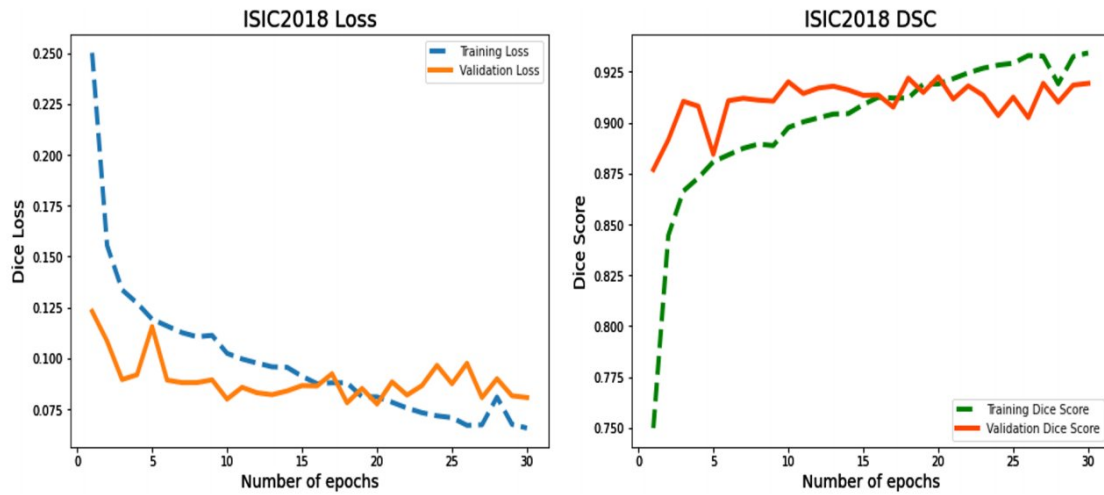


Figure 4.4: Progression of Training and Validation Dice Loss and Dice Coefficients over the number of epochs for ISIC 2018 dataset

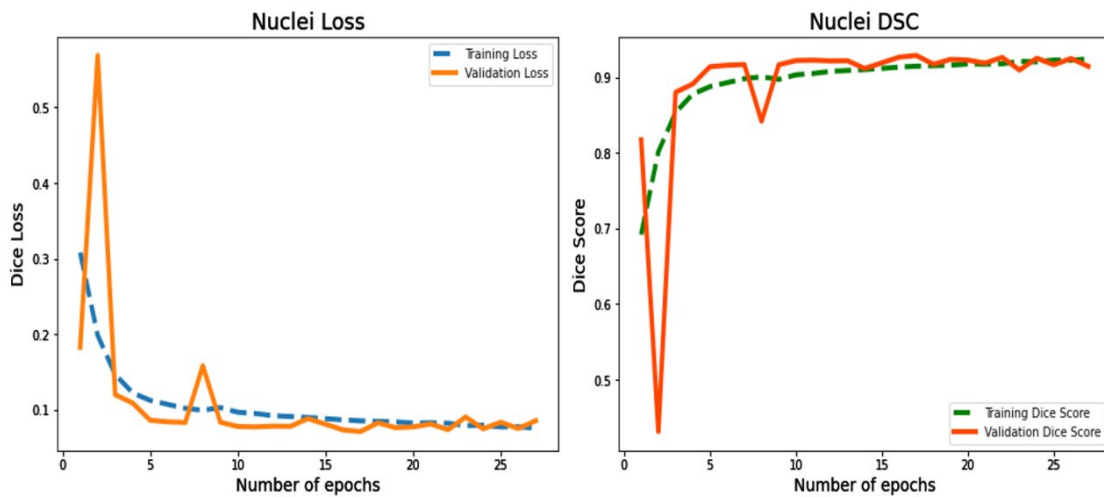


Figure 4.6: Progression of Training and Validation Dice Loss and Dice Coefficients over the number of epochs for 2018 Data Science Bowl dataset

4.4.3 Comparative analysis with current state-of-the-art architectures

We evaluated our proposed architecture on three different datasets and compared our results with that of the state-of-the-art models on these datasets. Table 4.5 shows the performance of the current state-of-the-art models on the ISIC 2018 dataset based on the Dice Coefficient. We can see that even though the base-

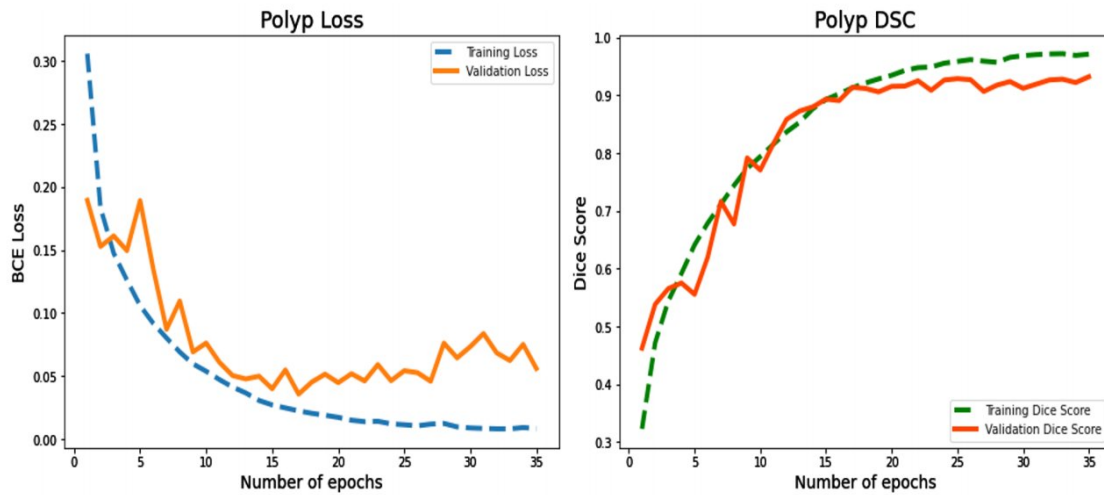


Figure 4.5: Progression of Training and Validation Dice Loss and Dice Coefficients over the number of epochs for CVC-ClinicDB dataset

line Double U-Net requires a huge number of data augmentations, it still ranks 6th on the table. The authors used around **50,000** training images to reach the given DSC score. The architectures of DS-TransUNet ranked 4th and Boundary Aware Transformer (BAT) ranked 5th are both based upon transformer architectures that are computationally very expensive to train due to their large number of parameters. Based on the encoder used, the number of parameters in these models can vary anywhere between **60 - 234 Million**. Ranked 3rd on this table is the RMSM U-Net which uses a complex dual attention mechanism that is incorporated on top of the traditional U-Net. The method ranked 1 in this table is a pre-processing technique that converts cartesian images to polar coordinate images using a center-point predictor network. Therefore, this method relies heavily on the pre-processing technique which requires a **model based pre-processing** approach both during training and inference. Now our model improves upon all the existing models discussed by using attention mechanism with a far less number of parameters (**36 Million**), and requires fewer images (**14,000**) to train with the help of a much cheaper pre-processing technique which is the color constancy algorithm.

To verify the robustness of our model, we further trained and tested our model on two other datasets. Table 4.6 shows the results of our proposed model on the

Rank	Model	DSC(%)	Year
1	Polar Res-U-Net++ [3]	92.53	2021
2	Our Model	91.68	2022
3	RMSM U-Net [13]	91.52	2021
4	DS-TransUNet [33]	91.32	2021
5	BAT [46]	91.20	2021
6	Double U-Net [29]	89.62	2020

Table 4.5: Comparative result of our model against the current state-of-the-art models on the ISIC 2018 dataset

CVC-ClinicDB dataset in comparison to the current state-of-the-art results. We can see that our model achieves the best DSC score (**94.35%**) which is higher than all the existing methods in this dataset. Similarly, our model achieves the best DSC score (**92.45%**) on the 2018 Data Science Bowl dataset as well. We can see the scores of our model compared to the state-of-the-art models in this dataset from table 4.7. From these experiments, we can successfully say that our model generalizes extremely well on segmentation tasks across various types of medical data.

Model	DSC(%)	IOU(%)	Precision(%)	Recall(%)
Our Model	94.35	89.32	97.37	87.50
Polar Res-U-Net++ [3]	93.74	89.77	94.88	93.68
MSRF-Net [43]	94.20	90.43	-	-
Double U-Net [29]	92.39	86.11	95.92	84.57
U-Net [41]	87.81	78.81	93.29	78.65

Table 4.6: Comparative result of our model against the current state-of-the-art models on the CVC Clinic-DB dataset

Model	DSC(%)	IOU(%)	Precision(%)	Recall(%)
Our Model	92.45	85.96	96.29	65.55
SSFormer-L [47]	92.30	86.14	-	-
MSRF-Net [43]	92.24	85.34	-	-
Double U-Net [29]	91.33	84.07	94.96	64.07
U-Net [41]	75.73	91.03	-	-

Table 4.7: Comparative result of our model against the current state-of-the-art models on the 2018 Data Science Bowl dataset

To summarize, if we compare the DSC scores of our proposed approach with that of the standalone Double U-Net, we can observe a clear distinction in results across all the datasets involved with a net improvement of **1.96%**, **2.06%** and **1.12%** over the lesion, polyp and nuclei datasets respectively.

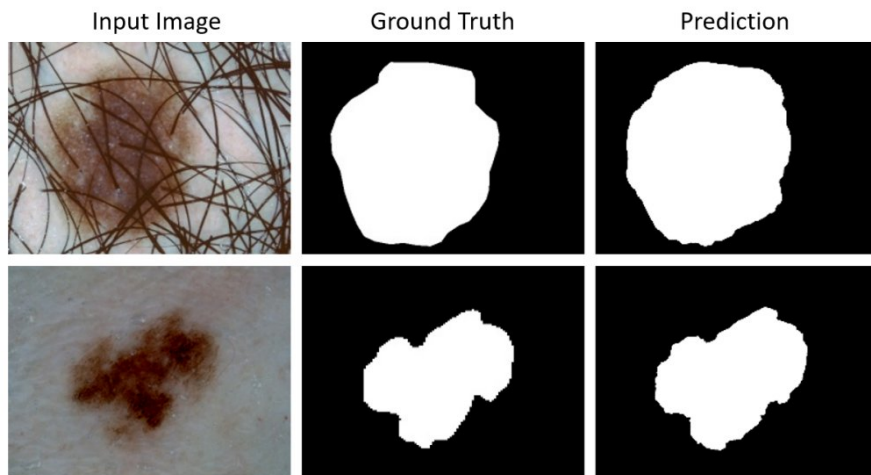
Modality	Double U-Net (DSC)(%)	Ours (DSC)(%)	Net improvement(%)
Polyp	92.39	94.35	1.96
Skin Lesion	89.62	91.68	2.06
Nuclei	91.33	92.45	1.12

Table 4.8: Overall improvement on the three benchmark datasets over standalone Double U-Net

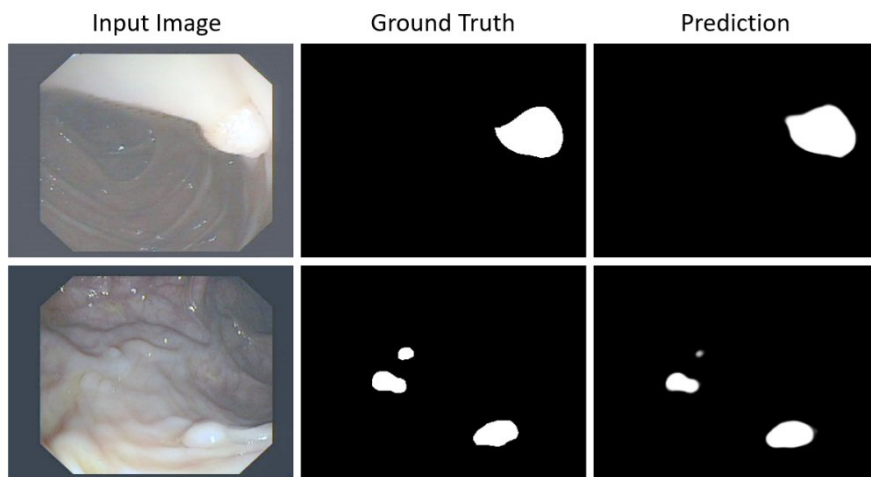
4.5 Qualitative Results

Figure 4.7 illustrates the outputs of our model for the three different datasets. We can observe that our model was able to segment out even flat polyp structures shown in Figure 4.7b and the application of CC helped to reduce the reflectance in each image. From Figure 4.7a we can observe that the variable shapes and sizes of skin lesions along with hair artefacts were also not a challenge as our model still

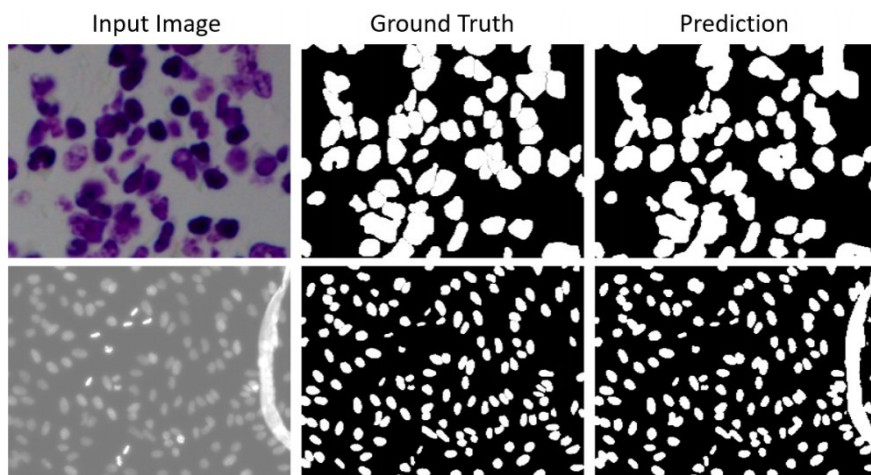
produced high quality segmentation masks for them. Moreover, our model was able to correctly segment out the densely connected nuclei cells as seen in the input images in 4.7c, irrespective of the color of the image. From this qualitative analysis we can conclude that our model mitigates the research challenges mentioned at the beginning of this report and generates good quality segmentation masks across all datasets 4.7a.



(a) Qualitative predictions of our model on skin lesion dataset



(b) Qualitative predictions of our model on polyp dataset



(c) Qualitative predictions of our model on nuclei dataset

Figure 4.7: Qualitative results of our model across three datasets

CHAPTER 5

CONCLUSION AND FUTURE WORK

In this paper, we propose a novel attention-based residual Double U-Net architecture for the task of medical image segmentation. Our approach incorporates three components over the standalone Double U-Net architecture. One is the addition of attention gates to the skip connections. We also added residual connections to the convolutional blocks and applied the color constancy algorithm to each dataset as a pre-processing step. As seen from the experiments our approach performs better than standalone Double U-Net across several datasets highlighting the robustness of our model. We achieved state-of-the-art DSC values in the polyp and nuclei datasets while having DSC score close to the state-of-the-art in the lesion dataset. For further improvements, we aim to design a simplified architecture to reduce the number of parameters while retaining similar accuracy in order to train the model faster. We can also look toward integrating different types of CNN blocks into our model with different hyperparameters.

BIBLIOGRAPHY

- [1] Mohammed A. Al-masni, Mugahed A. Al-antari, Mun-Taek Choi, Seung-Moo Han, and Tae-Seong Kim. Skin lesion segmentation in dermoscopy images via deep full resolution convolutional networks. *Comput. Methods Programs Biomed.*, 162:221–231, 2018.
- [2] Lucia Ballerini, Robert B Fisher, Ben Aldridge, and Jonathan Rees. A color and texture based hierarchical k-nn approach to the classification of non-melanoma skin lesions. In *Color medical image analysis*, pages 63–86. Springer, 2013.
- [3] Marin Bencevic, Irena Galic, Marija Habijan, and Danilo Babin. Training on polar image transformations improves biomedical image segmentation. *IEEE Access*, 9:133365–133375, 2021.
- [4] Jorge Bernal, Francisco Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez de Miguel, and Fernando Vilariño. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Comput. Medical Imaging Graph.*, 43:99–111, 2015.
- [5] Matt Berseth. ISIC 2017 - skin lesion analysis towards melanoma detection. *CoRR*, abs/1703.00523, 2017.

- [6] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. Alumentations: Fast and flexible image augmentations. *Inf.*, 11(2):125, 2020.
- [7] Juan C Caicedo, Allen Goodman, Kyle W Karhohs, Beth A Cimini, Jeanelle Ackerman, Marzieh Haghghi, CherKeng Heng, Tim Becker, Minh Doan, Claire McQuin, et al. Nucleus segmentation across imaging experiments: the 2018 data science bowl. *Nature methods*, 16(12):1247–1253, 2019.
- [8] Albert Cardona, Stephan Saalfeld, Stephan Preibisch, Benjamin Schmid, Anchi Cheng, Jim Pulokas, Pavel Tomancak, and Volker Hartenstein. An integrated micro-and macroarchitectural analysis of the drosophila brain by computer-assisted serial section electron microscopy. *PLoS biology*, 8(10):e1000502, 2010.
- [9] M. Emre Celebi, Wenzhao Guo, Y. Alp Aslandogan, and Paul R. Bergstresser. Skin lesion segmentation using clustering techniques. In Ingrid Russell and Zdravko Markov, editors, *Proceedings of the Eighteenth International Florida Artificial Intelligence Research Society Conference, Clearwater Beach, Florida, USA*, pages 364–369. AAAI Press, 2005.
- [10] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *CoRR*, abs/2102.04306, 2021.
- [11] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(4):834–848, 2018.
- [12] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th*

- European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII*, volume 11211 of *Lecture Notes in Computer Science*, pages 833–851. Springer, 2018.
- [13] G. Jignesh Chowdary, G. V. S. N. Durga Yathisha, Suganya G, and Premalatha M. Exploring dual-attention mechanism with multi-scale feature extraction scheme for skin lesion segmentation. *CoRR*, abs/2111.08708, 2021.
- [14] Kenneth W. Clark, Bruce A. Vendt, Kirk E. Smith, John B. Freymann, Justin S. Kirby, Paul Koppel, Stephen M. Moore, Stanley R. Phillips, David R. Maffitt, Michael Pringle, Lawrence Tarbox, and Fred W. Prior. The cancer imaging archive (TCIA): maintaining and operating a public information repository. *J. Digit. Imaging*, 26(6):1045–1057, 2013.
- [15] Noel C. F. Codella, David A. Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin K. Mishra, Harald Kittler, and Allan Halpern. 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 *15th IEEE International Symposium on Biomedical Imaging, ISBI 2018, Washington, DC, USA, April 4-7, 2018*, pages 168–172. IEEE, 2018.
- [16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society, 2009.
- [17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th*

- International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.
- [18] Timothy Dozat. Incorporating nesterov momentum into adam. In *International Conference on Learning Representations, Caribe Hilton, San Juan, Puerto Rico, May 2 - 4, 2016*, 2016.
- [19] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. In Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz, editors, *Medical Image Computing and Computer Assisted Intervention - MICCAI 2020 - 23rd International Conference, Lima, Peru, October 4-8, 2020, Proceedings, Part VI*, volume 12266 of *Lecture Notes in Computer Science*, pages 263–273. Springer, 2020.
- [20] Graham D. Finlayson and Elisabetta Trezzi. Shades of gray and colour constancy. In *The Twelfth Color Imaging Conference: Color Science and Engineering Systems, Technologies, Applications, CIC 2004, Scottsdale, Arizona, USA, November 9-12, 2004*, pages 37–41. IS&T - The Society for Imaging Science and Technology, 2004.
- [21] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial networks. *CoRR*, abs/1406.2661, 2014.
- [22] Manu Goyal, Moi Hoon Yap, and Saeed Hassanpour. Multi-class semantic segmentation of skin lesions via fully convolutional networks. In Elisabetta De Maria, Ana L. N. Fred, and Hugo Gamboa, editors, *Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2020) - Volume 3: BIOINFORMATICS, Valletta, Malta, February 24-26, 2020*, pages 290–295. SCITEPRESS, 2020.

- [23] Sebastian Gross, Manuel Kennel, Thomas Stehle, Jonas Wulff, Jens J. W. Tischendorf, Christian Trautwein, and Til Aach. Polyp segmentation in NBI colonoscopy. In Hans-Peter Meinzer, Thomas Martin Deserno, Heinz Handels, and Thomas Tolxdorff, editors, *Bildverarbeitung für die Medizin 2009: Algorithmen - Systeme - Anwendungen, Proceedings des Workshops vom 22. bis 25. März 2009 in Heidelberg*, Informatik Aktuell, pages 252–256. Springer, 2009.
- [24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016.
- [25] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 7132–7141. Computer Vision Foundation / IEEE Computer Society, 2018.
- [26] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5967–5976. IEEE Computer Society, 2017.
- [27] Saeed Izadi, Zahra Mirikharaji, Jeremy Kawahara, and Ghassan Hamarneh. Generative adversarial networks to segment skin lesions. In *15th IEEE International Symposium on Biomedical Imaging, ISBI 2018, Washington, DC, USA, April 4-7, 2018*, pages 881–884. IEEE, 2018.
- [28] Viren Jain, Benjamin Bollmann, Mark Richardson, Daniel R. Berger, Moritz Helmstaedter, Kevin L. Briggman, Winfried Denk, Jared B. Bowden, John M. Mendenhall, Wickliffe C. Abraham, Kristen M. Harris, Narayanan Kasthuri, Ken J. Hayworth, Richard Schalek, Juan Carlos Tapia, Jeff W. Lichtman, and H. Sebastian Seung. Boundary learning by optimization with topological

- constraints. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*, pages 2488–2495. IEEE Computer Society, 2010.
- [29] Debesh Jha, Michael A. Riegler, Dag Johansen, Pål Halvorsen, and Håvard D. Johansen. Doubleu-net: A deep convolutional neural network for medical image segmentation. In Alba García Seco de Herrera, Alejandro Rodríguez González, K. C. Santosh, Zelalem Temesgen, Bridget Kane, and Paolo Soda, editors, *33rd IEEE International Symposium on Computer-Based Medical Systems, CBMS 2020, Rochester, MN, USA, July 28-30, 2020*, pages 558–564. IEEE, 2020.
- [30] Debesh Jha, Pia H. Smedsrud, Michael A. Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D. Johansen. Kvasir-seg: A segmented polyp dataset. In Yong Man Ro, Wen-Huang Cheng, Junmo Kim, Wei-Ta Chu, Peng Cui, Jung-Woo Choi, Min-Chun Hu, and Wesley De Neve, editors, *MultiMedia Modeling - 26th International Conference, MMM 2020, Daejeon, South Korea, January 5-8, 2020, Proceedings, Part II*, volume 11962 of *Lecture Notes in Computer Science*, pages 451–462. Springer, 2020.
- [31] Debesh Jha, Pia H. Smedsrud, Michael A. Riegler, Dag Johansen, Thomas de Lange, Pål Halvorsen, and Håvard D. Johansen. Resunet++: An advanced architecture for medical image segmentation. In *IEEE International Symposium on Multimedia, ISM 2019, San Diego, CA, USA, December 9-11, 2019*, pages 225–230. IEEE, 2019.
- [32] Taehun Kim, Hyemin Lee, and Daijin Kim. Uacanet: Uncertainty augmented context attention for polyp segmentation. In Heng Tao Shen, Yueting Zhuang, John R. Smith, Yang Yang, Pablo Cesar, Florian Metze, and Balakrishnan Prabhakaran, editors, *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 2167–2175. ACM, 2021.

- [33] Ailiang Lin, Bingzhi Chen, Jiayu Xu, Zheng Zhang, and Guangming Lu. Ds-transunet: Dual swin transformer u-net for medical image segmentation. *CoRR*, abs/2106.06716, 2021.
- [34] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Anal.*, 42:60–88, 2017.
- [35] Martin Maska, Vladimír Ulman, David Svoboda, Pavel Matula, Petr Matula, Cristina Ederra, Ainhoa Urbiola, Tomás España, Subramanian Venkatesan, Deepak M. W. Balak, Pavel Karas, Tereza Bolcková, Markéta Streitová, Craig Carthel, Stefano Coraluppi, Nathalie Harder, Karl Rohr, Klas E. G. Magnusson, Joakim Jaldén, Helen M. Blau, Oleh Dzyubachyk, Pavel Krížek, Guy M. Hagen, David Pastor-Escuredo, Daniel Jimenez-Carretero, María J. Ledesma-Carbayo, Arrate Muñoz-Barrutia, Erik Meijering, Michal Kozubek, and Carlos Ortiz-de-Solorzano. A benchmark for comparison of cell tracking algorithms. *Bioinform.*, 30(11):1609–1617, 2014.
- [36] Sinan Naji, Hamid Abdullah Jalab, and Sameem Abdul Kareem. A survey on skin detection in colored images. *Artif. Intell. Rev.*, 52(2):1041–1087, 2019.
- [37] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII*, volume 9912 of *Lecture Notes in Computer Science*, pages 483–499. Springer, 2016.
- [38] Jiahua Ng, Manu Goyal, Brett Hewitt, and Moi Hoon Yap. The effect of color constancy algorithms on semantic segmentation of skin lesions. In Barjor Gimi and Andrzej Król, editors, *Medical Imaging 2019: Biomedical Applica-*

- tions in Molecular, Structural, and Functional Imaging, San Diego, California, United States, 16-21 February 2019*, volume 10953 of *SPIE Proceedings*, page 109530R. SPIE, 2019.
- [39] Ozan Oktay, Jo Schlemper, Loïc Le Folgoc, Matthew C. H. Lee, Mattias P. Heinrich, Kazunari Misawa, Kensaku Mori, Steven G. McDonagh, Nils Y. Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas. *CoRR*, abs/1804.03999, 2018.
- [40] Howard W Rogers, Martin A Weinstock, Steven R Feldman, and Brett M Coldiron. Incidence estimate of nonmelanoma skin cancer (keratinocyte carcinomas) in the us population, 2012. *JAMA dermatology*, 151(10):1081–1086, 2015.
- [41] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells III, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, volume 9351 of *Lecture Notes in Computer Science*, pages 234–241. Springer, 2015.
- [42] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.
- [43] Abhishek Srivastava, Debesh Jha, Sukalpa Chanda, Umapada Pal, Håvard D. Johansen, Dag Johansen, Michael A. Riegler, Sharib Ali, and Pål Halvorsen. Msrf-net: A multi-scale residual fusion network for biomedical image segmentation. *CoRR*, abs/2105.07451, 2021.
- [44] Jian-Qin Tang, Xiao-Yang Hou, Chun-Sheng Yang, Ya-Xi Li, Yong Xin, Wen-Wen Guo, Zhi-Ping Wei, Yan-Qun Liu, and Guan Jiang. Recent developments

- in nanomedicine for melanoma treatment. *International journal of cancer*, 141(4):646–653, 2017.
- [45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.
- [46] Jiacheng Wang, Lan Wei, Liansheng Wang, Qichao Zhou, Lei Zhu, and Jing Qin. Boundary-aware transformers for skin lesion segmentation. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention - MICCAI 2021 - 24th International Conference, Strasbourg, France, September 27 - October 1, 2021, Proceedings, Part I*, volume 12901 of *Lecture Notes in Computer Science*, pages 206–216. Springer, 2021.
- [47] Jinfeng Wang, Qiming Huang, Feilong Tang, Jia Meng, Jionglong Su, and Sifan Song. Stepwise feature fusion: Local guides global. *CoRR*, abs/2203.03635, 2022.
- [48] Zenghui Wei, Feng Shi, Hong Song, Weixing Ji, and Guanghui Han. Attentive boundary aware network for multi-scale skin lesion segmentation with adversarial training. *Multim. Tools Appl.*, 79(37-38):27115–27136, 2020.
- [49] Alexander Wong, Jacob Scharcanski, and Paul W. Fieguth. Automatic skin lesion segmentation via iterative stochastic region merging. *IEEE Trans. Inf. Technol. Biomed.*, 15(6):929–936, 2011.
- [50] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In Yoshua Bengio and Yann LeCun, editors, *4th International*

Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings, 2016.

- [51] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In Danail Stoyanov, Zeike Taylor, Gustavo Carneiro, Tanveer F. Syeda-Mahmood, Anne L. Martel, Lena Maier-Hein, João Manuel R. S. Tavares, Andrew P. Bradley, João Paulo Papa, Vasileios Belagiannis, Jacinto C. Nascimento, Zhi Lu, Sailesh Conjeti, Mehdi Moradi, Hayit Greenspan, and Anant Madabhushi, editors, *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*, volume 11045 of *Lecture Notes in Computer Science*, pages 3–11. Springer, 2018.