Advancing Medical Image Segmentation through Transformer-Enhanced Algorithms and Dataset Integration

By

Ahmed Suleman Salim

Fall-2021-MS-EE 363789 SEECS

Supervisor

Dr Arbab Latif

A thesis submitted in partial fulfillment of the requirements for the degree of Masters of Science in Electrical Engineering with Specialization in Artificial Intelligence and Autonomous Systems (MS EE AI & AS)

School of Electrical Engineering & Computer Science (SEECS)
National University of Sciences and Technology (NUST)
Islamabad, Pakistan
(Decemeber 2024)

Approval

It is certified that the contents and form of the thesis entitled "Advancing Medical Image Segmentation through Transformer-Enhanced Algorithms and Dataset Integration" submitted by Ahmed Suleman Salim have been found satisfactory for the requirement of the degree

Advisor : Dr. Arbab Latif

Signature: Latif

Date: 05-Nov-2024

Committee Member 1:Dr. Wajid Mumtaz

Signature:

28-Oct-2024

Committee Member 2:Dr. Muhammad Shahzad Younis

Signature:

Date: 05-Nov-2024

Signature: ______________________

Date: ______________________

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis entitled "Advancing Medical Image Segmentation through Transformer-Enhanced Algorithms and Dataset Integration" written by Ahmed Suleman Salim, (Registration No 00000363789), of SEECS has been vetted by the undersigned, found complete in all respects as per NUST Statutes/Regulations, is free of plagiarism, errors and mistakes and is accepted as partial fulfillment for award of MS/MPhil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

Signature: Latif

Name of Advisor: Dr. Arbab Latif

Date: 05-Nov-2024

HoD/Associate Dean:

Date: 05-Nov-2024

Signature (Dean/Principal):

Date: 05-Nov-2024

AUTHOR'S DECLARATION

I hereby declare that this submission titled "Advancing Medical Image Segmentation through Transformer-Enhanced Algorithms and Dataset Integration" is my own work. To the best of my knowledge it contains no materials previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any degree or diploma at NUST SEECS or at any other educational institute, except where due acknowledgement has been made in the thesis. Any contribution made to the research by others, with whom I have worked at NUST SEECS or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except for the assistance from others in the project's design and conception or in style, presentation and linguistics, which has been acknowledged. I also verified the originality of contents through plagiarism software.

Student Name:Ahmed Suleman Salim

Student Signature: _______________

Date: 05-Nov-2024

Certificate for Plagiarism

It is certified that PhD/M.Phil/MS Thesis Titled "Advancing Medical Image Segmentation through Transformer-Enhanced Algorithms and Dataset Integration" by Ahmed Suleman Salim has been examined by us. We undertake the follows:

a. Thesis has significant new work/knowledge as compared already published or are under consideration to be published elsewhere. No sentence, equation, diagram, table, paragraph or section has been copied verbatim from previous work unless it is placed under quotation marks and duly referenced.

b. The work presented is original and own work of the author (i.e. there is no plagiarism). No ideas, processes, results or words of others have been presented as Author own work.

c. There is no fabrication of data or results which have been compiled/analyzed.

d. There is no falsification by manipulating research materials, equipment or processes, or changing or omitting data or results such that the research is not accurately represented in the research record.

e. The thesis has been checked using TURNITIN (copy of originality report attached) and found within limits as per HEC plagiarism Policy and instructions issued from time to time.

Name & Signature of Supervisor

Dr. Arbab Latif

Signature : Latif

FORM TH-4

National University of Sciences & Technology

MASTER THESIS WORK

We hereby recommend that the dissertation prepared under our supervision by: (Student Name & Reg. #) Ahmed Suleman Salim [00000363789]

Titled: Advancing Medical Image Segmentation through Transformer-Enhanced Algorithms and Dataset Integration

be accepted in partial fulfillment of the requirements for the award of Master of Science (Electrical Engineering) degree.

Examination Committee Members

1. Name: Wajid Mumtaz Signature: 03-Dec-2024 11:27 PM

2. Name: Muhammad Shahzad Younis Signature: 03-Dec-2024 11:27 PM

Supervisor's name: Arbab Latif Signature: 03-Dec-2024 11:55 PM

Salman Abdul Ghafoor
HoD / Associate Dean

11-December-2024
Date

COUNTERSINGED

11-December-2024
Date

Muhammad Ajmal Khan
Principal

THIS FORM IS DIGITALLY SIGNED

DEDICATION

To my dear family, your unwavering support, both emotionally and financially, has been the cornerstone of my academic journey. Mom, your strength and encouragement have been my guiding light, propelling me forward even in the face of challenges. This research work is a testament to your belief in me, and I dedicate it to you all with immense gratitude and love. Thank you for being my rock and my greatest source of inspiration.

Ahmed Suleman Salim

ACKNOWLEDGEMENTS

After Allah Almighty, I would like to thank my supervisor, for his support and encouragement. His profound belief in my skills and capabilities helped me to complete this thesis. His positive feedback and constructive criticism have pushed me to do my best, and I am glad for the opportunity to work under his supervision. Lastly, I would like to show my gratitude to my committee members. they have been equally supportive and helpful throughout this research period and guided me to achieve the desired goal.

Ahmed Suleman Salim

Contents

List of Tables

List of Figures

List of Abbreviations and Symbols

Abbreviations

TFA - Transformer Attention

ANA - Attention In Atention

DDTI - Digital Database Thyroid Image

BUSI - Breast Ultrasound Images

ISIC - International Skin Imaging Collaboration

CNN - Convolutional Neural Networks

CAD - Computer Aided Diagnosis

CNN - Convolution Neural Network

FCN - Fully Convolution Network

Abstract

In an era marked by a rising prevalence of health issues, the significance of a reliable and efficient system for disease detection is a need. With the successful integration of Transformer models in computer vision, researchers are increasingly delving into their application in medical image segmentation. Particularly, there's a growing exploration of combining Transformers with convolutional neural networks featuring coding-decoding architectures. The fusion has demonstrated remarkable achievements in medical image segmentation. In this research, the main goal is to create advanced algorithms that can match or even surpass the accuracies achieved by currently established models when applied to particular datasets. Involving pushing the boundaries of existing methodologies and techniques to enhance the performance of the segmentation process in medical imaging. The focus will be on innovating novel approaches that can handle various challenges present in medical image segmentation tasks, such as noise, variability in anatomy, and imaging modalities. By developing state-of-the-art algorithms, the aim is to contribute to the advancement of the field and potentially improve diagnostic and analytical capabilities in clinical settings. By incorporating diverse datasets representative of different medical conditions, this research attempts to enhance the effectiveness and generalizability of our findings. The aim is to enhance medical image segmentation techniques and to develop robust algorithms and methodologies for accurate medical image analysis and segmentation.

Chapter 1

Introduction and Motivation

1.1 Background

Image segmentation is an important component of quantitative medical imaging analysis, often serving as the initial step for examining anatomical structures [1]. Since the advancements in deep learning, Fully Convolutional Neural Networks (FCNNs), particularly "U-shaped" encoder-decoder architectures [2, 3, 4], have delivered state-of-the-art performance in various medical segmentation tasks [5, 6, 7] . In the classic U-Net architecture, the encoder's role is to learn global contextual representations by progressively down sampling the features extracted [1]. Conversely, the decoder side up samples these features back to the input resolution to achieve pixel wise semantic prediction. Additionally, skip connections integrate the encoder's output with the decoder at different resolutions, helping to restore spatial information lost during down sampling.

Although FCNN based approaches exhibit strong representation of learning capabilities, their ability to learn long-range dependencies is constrained by their localized receptive fields [8, 9]. This constraint in acquiring multi-scale information leads to less effective segmentation of structures with diverse shapes and sizes. To address this issue, some studies have employed atrous convolutional layers to expand the receptive fields [10, 11]. However, the inherent locality of convolutional layers still restricts their learning capacity to relatively small regions. To enhance non-local modeling capability, integrating self-attention modules with convolutional layers has been proposed [12, 13].

The Fully Convolutional Neural Network (FCN) [14] was among the pioneering deep learing models used to aid image segmentation. Later more advancement was made to this architecture to create U-Net, which achieved notable segmentation results by leveraging the availability of large training datasets [15].U-Net's architecture is built on an encoding and decoding pathway. In the encoding pathway, numerous feature maps are derived from the input data, with dimensionality progressively reduced.. The decoding path then produces segmentation maps (matching the input size) through up-convolutions.Various adaptations of U-Net have been introduced [16, 17], with many focusing on changes to the skip connections. In certain U-Net extensions, the feature maps within skip connections go through an additional processing phase, such as attention gates [17], before concatenation. A key limitation of these designs is that the processing step is performed independently on each set of feature maps, which are simply concatenated afterward.

Following the development of U-shaped architectures, models like ResUNet [18] and Attention R2UNet [19] emerged. However, these architectures frequently face challenges in capturing and leveraging multi-scale contextual features within a single stage, which is critical in medical imaging, where the target region often closely resembles surrounding areas. Integrating multi-scale contextual information helps provide context around the target and reduce ambiguity in decision-making [20]. Recent methods, including DeepLabV3 [21], PoolNet [22], PSPNet [23], and CE-Net [24], have been designed to incorporate multi-scale features. Yet, these methods primarily emphasize high-level feature extraction, often overlooking the valuable spatial details found in lower-level features.

Methods based on CNN are effective at feature extraction yet struggle with capturing long ranged dependencies which is because of limitations associated with convolution process. As a result, these methods may perform sub optimally when targeting regions with significant variations in texture, size, and shape. To address this, some researchers have incorporated attention mechanisms with CNNs [25, 26, 27]. Moreover, the success of Transformers in computer vision has opened up additional approaches [28, 29]. Transformers, operating as sequence to sequence models, avoid convolution and instead utilize self-attention mechanisms to derive image feature information and capture long-range relationships.

Transformers have surpassed state of the art performance in various tasks related to vision . While they are god at modelling global context of input image, Transformers have limitations when capturing fine grained details, especially in medical images, due to a lack of spatial induction bias when modelling local information. Moreover, Transformer-based network structures generally require large datasets to perform effectively [28]. This is where CNN architectures can compensate well, as they are more effective with smaller datasets.

Recent studies have investigated the integration of CNNs and Transformers for medical segmentation tasks. Approaches like TransUNet [30] and related models [31, 32, 33] employ CNNs as foundational networks, while Transformers are utilized to capture long range dependencies within the high level feature representations. However, these methods often fail to find the rich spatial details present in the shallow layers, focusing instead on single-scale context modeling and failing to address cross-scale dependencies and consistency. Furthermore, [34] suggests that employing only one or two Transformer layers may not be sufficient to effectively merge the long-range features extracted by CNNs.

1.2 Research Problem

U-Nets have demonstrated powerful performance in medical image segmentation but are often limited in capturing global context due to their reliance on convolutional operations. Vision Transformers offer improved performance through self-attention mechanisms but are susceptible to overfitting. Hybrid models that combine Transformers and U-Nets present a promising innovation in medical image segmentation. However, challenges remain in optimizing these hybrid architectures to effectively leverage both local and global contextual information while managing computational complexity.

Existing segmentation methods also had low segmentation accuracy on medical images since there is a large variance in noise, artifacts and contrast of medical images. These imperfections might be sensitive to current solutions thereby affecting their robustness and reliability for clinical practices. Further, the specialized models trained and validated on specific datasets may not necessarily generalize well on new, unseen data. The diversity of medical images necessitates strong generalizable segmentation algorithms.

For binary class medical image segmentation tasks (e.g., presence or absence of a condition), complex and computationally heavy networks are commonly used. Nevertheless, deep networks are not always appropriate, effective, powerful, or necessary via diverse datasets especially in the case of accurate,true binary segmentation. Exploring and fine-tuning architecture with careful consideration of balance among model complexity, computational efficiency and high-quality segmentation performance.

1.3 Problem Statement

In modern healthcare, the accurate segmentation of medical images is of immense importance since accurate segmentation facilitates diagnosis, treatment planning, and disease monitoring. The deep learning techniques applied to medical image segmentation have seen continuous development, but we still need to find ways to enhance their effectiveness and efficiency. Moreover, there are many types of medical images, and corresponding representations are highly varied because of differences in shape, size and appearance, which compounds this challenge. Confronting the complexities are needed to improve clinical decision-making and ultimately patient outcomes.

1.4 Solution Statement

To address these challenges in medical image segmentation, this thesis proposes the development of advanced algorithms that integrate Transformer models with convolutional neural networks (CNNs) in an encoding-decoding architecture. The primary goal is to enhance the accuracy and versatility of segmentation across various medical conditions and imaging modalities, thereby alleviating workload and improving diagnostic accuracy.

Thus, this thesis be emphasis on a hybrid architecture, combining the Transformer model with a U-Net. The proposed architecture is a hybrid model that takes advantage of both Transformers (which are known to capture long-range dependencies through self-attention mechanisms) and U-Nets (which are strong in preserving fine spatial details in medical images). The model is then able to outperform the segmentation accuracy of classical CNN-based methods by merging these architectures. Through implementation, evaluation, and potential integration of this hybrid approach, the thesis can provide novel perspectives and set methods on medical image segmentation.

1.5 Thesis Organization

1.5.1 Research Objectives

The focus of this research is to contribute towards development of medical imaging within this field of image segmentation through new deep learning approaches that cater specifically for medical imaging. Image segmentation has become a necessity for precise diagnosis and therapy in many branches of medicine to determine areas in images which are the most significant and need detailed study. Therefore, this project aims at developing new deep learning structures and methods to increase the precision and performance of segmentation algorithms resulting in better diagnosis and patient care.

While investigating new deep learning architectures, this work will analyze the role of attention mechanisms as feature augmentation tools in image segmentation. Attention mechanisms have been very beneficial across a number of tasks in computer vision by enabling the model to concentrate on the pertinent aspects while disregarding a significant amount of irrelevant information. Therefore, this research aims at employing attention mechanisms to the segmentation processes to enhance feature quality and consequently the segmentation accuracy. Particular emphasis will be placed on the effectiveness of self-attention and spatial attention mechanisms in providing focus on salient features in medical images.

Additionally, the research involves a performance comparison with the existing segmentation techniques in order to test the proposed deep learning methods. The purpose of the study is to assess the new techniques' segmentation results against the best available methodologies, demonstrating the defects and advantages of the novel methods, thus pointing to the directions where more work is needed. It is essential to perform Out-of-Distribution (OOD) data tests to assess the generalization properties of the developed models. The outcome of this evaluation should have implications in relation to their robustness and the extent to which the models can cope in real-life situations.

Lastly, the research aims to ensure model adaptability through domain shift training and testing, maintaining consistent performance across diverse data domains and clinical environments. Through comprehensive evaluation methodologies, this study seeks to validate the effectiveness and applicability of the developed image segmentation techniques in clinical practice. The main objectives of the thesis are outlined below:

- Develop a novel deep learning model for enhanced image segmentation for medical imaging.
- Optimize encoder-decoder block interconnections to preserve feature integrity across various scales.
- Create attention mechanisms as feature enhancement tools to improve the features quality in the skip connections of the model.
- Enhance bottleneck features using transformer-based attention for superior segmentation performance.
- Validate the robustness and effectiveness of the model in medical image segmentation through thorough comparisons with current methodologies.

1.5.2 Research Contributions

This work introduces an innovative network structure, referred to as the proposed model, designed for medical image segmentation. The model utilizes a well-established encoder-decoder framework and integrates several important improvements to enhance the extraction of relevant features. These improvements consist of a booster architecture, local-global feature enhancement, skip connections based on normalized focal modulation, and transformer-based attention mechanisms at the bottleneck. The primary motivation is to enhance the extraction of critical features, hence improving the accuracy and effectiveness of medical segmentation tasks.

Proposed model features a backbone made up of CNNs with two branches in parallel with a booster architecture integrated into it. While the multiscale feature information is extracted using CNNs, the booster simultaneously captures global context of the information to model longe range dependencies. By effectively maximizing the retention of spatial information within low-level semantic features—due to their lower computational cost and importance—the model seeks to enhance computational efficiency without sacrificing segmentation quality.

In the decoding part, the encoder structure is reused, and a Transformer Attention (TFA) mechanism is incorporated at skip connection from the encoder booster to the decoder booster. Retrieval of both global and local information in the decoding stage is enhanced due to addition of TFA in skip connections. Moreover, TFA, combined with Attention In Attention (ANA), is utilized at the bottleneck to strengthen the connections between the decoder blocks, creating dense links which help feature retention throughout the up sampling process.

Main contributions of this thesis can be summarized as follows:

1. **Novel Architecture :** The proposed model utilizes a parallel booster design for both the encoder and decoder, due to which versatile feature sets are extracted improving segmentation performance.
2. **Dense Interconnections :** Strong interconnections between decoder blocks, establishing dense links that preserve improved features during the crucial upsampling process, contributing to maintaining feature integrity across different scales.
3. **Enhanced Feature Information:** By applying TFA attention at the skip connections, the model improves its ability to capture contextual information and intricate details .
4. **Transformer-Based Attention:** Strategically integrated at the bottleneck, enhances feature representation. This method, along with improvements to local and global characteristics, guarantees that crucial information must be preserved and effectively utilized throughout the segmentation process.
5. **Validation and Comparison:** Robustness and generalizability of the proposed model are validated through comprehensive comparisons with current state-of-the-art methods. This analysis highlights the model's efficacy and competitive performance in medical image segmentation.

1.5.3 Overview

Throughout this introductory chapter, we have highlighted the pressing need for improved methodologies in field of image segmentation. In the upcoming chapters, we will carefully navigate the landscape of research in this domain, with each section serving as a crucial step toward our overarching goal.

Chapter 2 will provide a comprehensive literature review, tracing the evolution of methodologies and insights gained by the research community over time. Building on this foundational understanding, Chapter 3 will introduce our meticulously crafted methodology, designed to tackle the complex challenges associated with image segmentation. This will naturally lead into Chapter 4, where we will present the empirical results obtained from our methodological approach, offering clear insights into their effectiveness and applicability.

In Chapter 5, we will engage in a thoughtful discussion that examines the implications of our findings, situating them within the broader context of image segmentation research. Chapter 6 will summarize the conclusions drawn from our study, synthesizing key insights and their implications. Finally, in Chapter 7, we will distill our findings into actionable recommendations aimed at informing clinical practice and guiding future research efforts.

Through this cohesive framework, our goal is not only to advance the field of medical image segmentation but also to empower both clinicians and researchers in their quest for improved diagnostic accuracy and patient care.

Chapter 2

Literature Review

2.1 Medical Image Segmentation

Medical image segmentation employs computer image processing techniques to analyze and process both 2D and 3D images. The goal of this process is to achieve segmentation, extraction, three-dimensional reconstruction, and visualization of human organs, soft tissues, and pathological areas. It divides an image into various regions based on the similarities or differences within those areas. This approach enables physicians to conduct qualitative and quantitative analyses of lesions and other areas of interest, greatly improving the accuracy and reliability of medical diagnoses. Currently, various tissues and organs are identified within the image cells as focal points of analysis.

Image segmentation has become a prominent topic in computer vision and image understanding research. It involves dividing an image into disjoint regions based on characteristics like grayscale, color, spatial texture, and geometric shapes. The goal is to ensure consistency or similarity within regions and distinct differences between them. Image segmentation can be categorized into semantic segmentation, instance segmentation, and panoramic segmentation, based on the level of granularity. Medical image segmentation is generally regarded as a semantic segmentation task. Currently, there is a growing number of research domains within image segmentation, including satellite image segmentation, medical image segmentation, and applications in autonomous driving [35, 36]." The continual development of new network structures has progressively improved segmentation methods, resulting in increasingly accurate outcomes. However, no single algorithm is universally applicable to all types of segmentation tasks.

Although traditional image segmentation techniques are simpler and quicker, they often do not match the accuracy of deep learning-based methods. Prominent traditional approaches include threshold-based segmentation [37], region-based segmentation [38], and edge detection-based segmentation [39]. These methods utilize concepts from digital image processing and mathematics for image segmentation but may lack precision in detail. While they offer fast calculation and segmentation speeds, they do not consistently ensure accuracy in the finer aspects of segmentation.

Deep learning has revolutionized image segmentation, with methods such as the fully convolutional network (FCN) leading the way. FCNs were among the first to successfully apply deep learning to image semantic segmentation. This pioneering work marked a significant advancement in using convolutional neural networks for this purpose. Deep learning based methods have since achieved remarkable accuracy, surpassing traditional segmentation techniques [40, 41, 42].

2.2 Deep Learning Architectures for Segmentation

The structure of segmentation networks in CNNs has evolved significantly. Initially, the modification involved replacing the last two fully connected layers in classification networks with convolutional layers. The foundation of medical image segmentation networks relies on deep structures like encoder-decoder architecture.

LeNet and AlexNet are foundational network models, both recognized as relatively shallow architectures. AlexNet, with a higher number of parameters than LeNet, introduced the innovative concept of applying a pooling layer after each convolutional layer, a technique that remains widely used today. VGG further advanced AlexNet by increasing the depth of the network, using several consecutive 3×3 convolutional filters instead of larger ones. This approach preserved the receptive field size while improving network depth and feature extraction. VGG's structure is notable for its straightforward design, featuring consistent convolution and pooling layer sizes, highlighting how deeper networks can enhance performance. However, increasing the network's depth can sometimes result in challenges like overfitting and vanishing gradient issues.

To address these issues, GoogleNet [43] introduced a modular approach with the Inception structure, increasing the network's depth and width while reducing the number of parameters.The Inception module utilizes convolutional filters of various sizes along with pooling layers, merging the outputs to create a network with a depth of 22 layers. CNN architectures have evolved from AlexNet's seven layers to VGG's 19 layers, and further to GoogleNet's 22 layers. However, beyond a certain depth, further increases do not always improve performance and can slow network convergence.

ResNet proposed by He et al. [44], a 152-layer network, to train deeper networks effectively. ResNet addresses depth-related issues with shortcuts comprising of residual blocks. In the each module in ResNet it composed of multiple layers and a shortcut that connects the module's input and output, adding them before ReLU activation. The resulting output is then passed through ReLU to generate the final output of the block [45].

The encoder-decoder architecture, which combines a CNN-based encoder with a decoder, forms the foundation of semantic segmentation networks. The encoder, often a CNN used for the classification tasks, help to extracts and compacts features from images, producing a feature map of low resolution. Decoder then maps this low resolution feature map to a high resolution pixel space, enabling pixel wise category labeling. SegNet [46] is a classic example of an encoder-decoder structure, with its encoder and decoder corresponding one-to-one in spatial size and number of channels. Innovations in semantic segmentation networks primarily focus on optimizing the encoder-decoder structure and improving efficiency, particularly the decoder's impact on the overall segmentation results.

2.3 FCN Based Segmentation

In the realm of CNN-based image segmentation, significant advancements have been made through successive iterations of models aimed at overcoming inherent challenges. The evolution began with introduction of (FCN), which replaced fully connected layers with convolutional layers to enable pixel-wise classification, addressing the limitation of one-dimensional classification output [47].

Building upon FCN, DeepLab v1 [48], refined segmentation by reducing pooling stride and padding size, introducing atrous convolution for larger receptive fields, and integrating Conditional Random Fields (CRF) for improved boundary delineation. DeepLab v2 [49] expanded on these concepts with atrous spatial pyramid pooling (ASPP), leveraging ResNet-101 for deeper feature extraction and eliminating downsampling to preserve spatial resolution.

DeepLab v3 [50] further enhanced multi-scale context capture through cascaded atrous convolution modules and ASPP, achieving notable improvements in segmentation accuracy over its predecessors. DeepLab v3+ [51] extended this progress by introducing a decoder module for finer segmentation along boundaries and employing the Xception model for enhanced speed and robustness. Each iteration addressed specific drawbacks, such as detail loss and scalability issues, resulting in more precise and context-aware segmentation models. These advancements have significantly contributed to the field of semantic segmentation, enhancing capabilities across various domains of image analysis and medical diagnostics.

2.4 UNet Based Segmentation

U-Net [52] is a CNN architecture designed around an encoder-decoder framework. The encoder, known as the contracting path, consists of successive convolutional layers with 3x3 filters followed by Rectified Linear Unit (ReLU) activations, forming what is termed a convolution block (conv-relu-conv-relu). To decrease the spatial dimensions of feature maps while progressively increasing the number of feature maps Max Pooling layers are used. This technique helps compress the information into a lower dimensional latent representation. The decoder, or expanding path, mirrors the encoder structure but replaces Max Pooling with up-convolution or transpose convolution to restore spatial dimensions. Here, the number of feature maps decreases progressively. This facilitates the passage of semantic information crucial for accurate segmentation. In this architecture, at the final layer, a softmax function is applied to classify the pixels in the segmentation map. The U-Net is said to have learned segmentation tasks efficiently because it is optimized end-to-end with a cross-entropy loss.

An improvement of the U-net architecture which adds recursively nested and densely connected skip pathways is U-Net++ [53]. Inspired by dense connections, these connections prevent the gradient cells from vanishing which results to better propagation of features across layers due to their reusability. The nested design also utilizes the encoders well by transferring adequate semantic information to the decoders which is important in performing medical image segmentation (MIS) accurately. U-Net++ also incorporates collecting segmentation maps around the intermediate layers to ensure deep supervision which enables learning features across different scales. In contrast to classical U-Net, this approach uses convolutional layers in skip pathways and Dense skip connections to facilitate gradient flow in the network while also using dense connections in the network to improve segmentation accuracy.

R2U-Net [54] is a variant of the U-Net architecture that combines residual and recurrent techniques to address the challenges of gradient propagation in deep convolutional networks. While deep networks are effective for many computer vision tasks, they often encounter issues such as vanishing gradients, which can hinder training. To mitigate this, R2U-Net introduces skip or identity connections, where outputs from previous blocks are directly fed into subsequent blocks, thus bypassing the current block operations. This concept draws inspiration from ResNets [55], which pioneered the use of such skip connections to enable training of very deep networks. Recurrent networks, traditionally used for sequential data like natural language and speech signals, also play a crucial role in R2U-Net. Unlike traditional U-Net's crop and copy approach, R2U-Net employs simpler feature concatenation from the encoder to the decoder. This modification maintains a similar parameter count while significantly enhancing performance in various segmentation tasks such as blood vessel, skin lesion, and lung lesion segmentation. Empirical studies demonstrate that R2U-Net consistently outperforms standard U-Net architectures in these applications.

2.5 Attention Unet Based Segmentation

Attention U-Net [56] introduces attention gates within its skip pathways, making it a hybrid architecture aimed at enhancing the precision of segmentation maps. These attention gates selectively pass crucial features to the decoder while suppressing redundant information. Initially employed in language tasks with great success, the concept of attention mechanisms has been adapted for visual tasks to improve the learning of local context. There are two types of attention gates: soft attention, which computes a weighted combination of inputs using differentiable functions with weights between 0 and 1, and hard attention, which makes discrete decisions. Attention U-Net utilizes soft attention gates in its skip pathways to enhance feature selection and focus on regions of interest (RoI), which is beneficial when datasets exhibit significant variations in RoI shapes and sizes. Additionally, residual and dense connections between the encoder and decoder components contribute to feature reuse and gradient flow, as observed in previous architectural enhancements. Although attention gates introduce additional computational overhead, they effectively elevate segmentation accuracy by highlighting salient features and refining the focus on relevant RoIs.

Furthermore, hybrid models like BCDU-Net [57] and CPF-Net [58] combine attention mechanisms with convolutional architectures to capture advanced temporal dependencies and inte- grate higher-level semantic data. While BCDU-Net achieves improved accuracy and robustness through the fusion of feature maps. Despite poten- tial drawbacks such as increased computational complexity and heavy parameters, these models represent promising avenues for enhancing medical image segmentation.

2.6 Transformer Based Segmentation

Trans U-Net [59] is an innovative architecture that integrates visual transformers (ViT) into the traditional U-Net framework to address the limitations of fully convolutional networks (FCNs) when handling significant variations in shape and size across datasets. Visual transformers have demonstrated remarkable success in sequence-to-sequence prediction tasks such as language and speech translation by leveraging multi-head self-attention mechanisms for capturing global dependencies. However, U-Net alone struggles with global spatial dependencies, while transformers may overlook low-level details crucial for precise segmentation. Trans U-Net combines the strengths of both architectures: it uses transformers to encode patch-wise image features into global representations, while maintaining local details with CNN-based feature extraction in the encoder. The encoded features from both CNN and transformer paths are appropriately concatenated before feeding into the decoder, which upsamples these features to the original image dimensions. This hybrid approach enhances localization accuracy of regions of interest (RoI) by aggregating features at multiple levels through skip connections. Consequently, Trans U-Net mitigates issues of under-segmentation and over-segmentation, improves global context awareness, and enhances semantic information extraction for more accurate image segmentation tasks.

Following the groundbreaking success of the Vision Transformer (ViT), there has been significant progress in vision-related tasks. DeiT [60], for example, concentrated on refining training methods for ViT architectures, which enhanced performance. The Pyramid Vision Transformer [61] which introduces a pyramid based architecture incorporating Shifted Relative Attention mechanisms, which lowered computational complexity without compromising performance. The Swin Transformer [62] made another important advancement by employing a window-based attention mechanism to enhance feature locality, addressing some of the limitations previously seen in transformer models. Transformers have also been adapted for specific tasks in computer vision. SETR (Semantic Segmentation Transformer) applies transformers to semantic segmentation tasks, with ViT serving as its core architecture. Xie et al. [63] developed SegFormer, which offers a simplified and efficient transformer-based model for semantic segmentation. Additionally, Wang et al. [64] proposed a U-shaped transformer known as Uformer, model designed for image restoration, demonstrating the flexibility and effectiveness of transformers across various applications in computer vision.

2.7 Hybrid Transformers and Unet based Segmentation

Transformers have become a highly effective tool in computer vision, especially in medical image segmentation, attracting significant interest from researchers due to their impressive performance. TransUNet [65] stands out as a pioneering approach in this field by integrating Transformer architecture with the traditional UNet encoder. This innovative fusion diverges from conventional methods by operating on high-level features, effectively capturing intricate spatial dependencies within medical images. By leveraging the hierarchical representations of UNet and the attention mechanisms of Transformers, TransUNet significantly enhances segmentation performance, setting a new standard in medical image analysis.

TransFuse [66], which introduces a novel perspective by concurrently bridging CNNs and Transformers. A BiFusion fusion module is fitted to the core of this approach, which combines shallow features from the encoders with features extracted by Transformers. This integrated approach enhances the overall understanding of the input data, utilizing both architectures to boost segmentation accuracy.

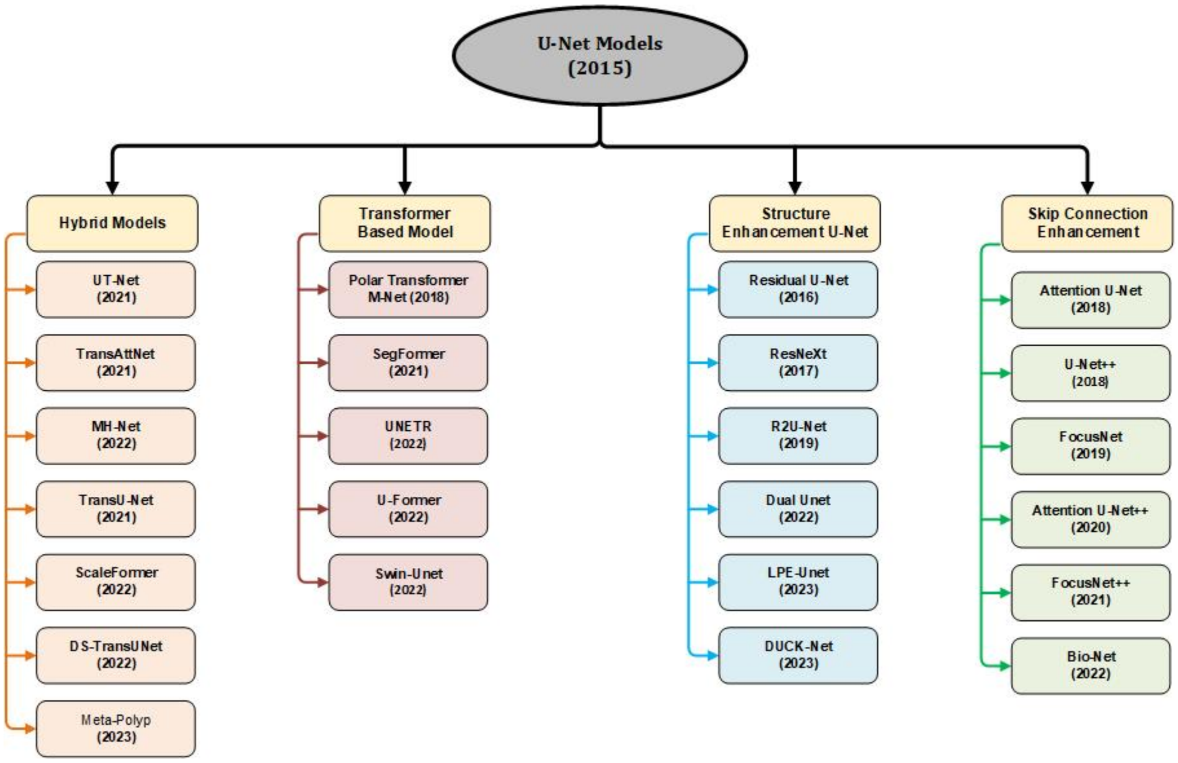

Figure 2.1: Literature Review Flowchart for Medical Segmentation.

On the other hand, TransAttunet [67] presents the Self-Aware Attention (SAA) module, which integrates Transformer Self-Attention (TSA) with Global Spatial Attention (GSA) to capture non-local interactions between encoder features.This enriches the segmentation process by enhancing feature integration across multiple scales. However, challenges remain in effectively amalgamating spatial and channel details crucial for precise segmentation tasks.

The integration of attention mechanisms has undoubtedly propelled advancements in medical image segmentation. Yet, early methodologies often face challenges in balancing spatial and channel information, potentially limiting segmentation precision. Future research endeavors will continue to explore hybrid CNN-Transformer architectures to optimize feature representation and further enhance the efficacy of medical image segmentation techniques.

2.8 Thyroid Nodule Segmentation

The thyroid gland, located in the anterior region of the neck beneath the thyroid cartilage, plays a crucial role in metabolic regulation by producing thyroid hormone [68]. Thyroid nodules are common and vary widely in their characteristics, ranging from well-defined to irregular shapes. These nodules can appear solid, cystic, or a combination of both. While approximately 4% to 7% of thyroid nodules are palpable, a much larger percentage, between 19% and 67%, is incidentally detected during ultrasound exams. Thyroid nodules are classified based on echogenicity into hypoechoic, isoechoic, or hyperechoic types. Studies suggest that hypoechoic nodules with irregular borders have a higher likelihood of being malignant. The incidence rate of malignant thyroid nodules is estimated to be between 0.1% to 0.2%.

Understanding the characteristics of thyroid nodules and their potential for malignancy is essential in clinical practice [69]. The variability in nodule presentation, ranging from well-defined shapes to irregular forms, highlights the need for comprehensive examination and diagnostic evaluation. Many people have nodules that're quite common, among the general populace; however the greater prevalence of nodules found during ultrasound scans is worth noting. Moreover analyzing thyroid nodules according to their echogenicity provides information about their characteristics, which can help with evaluating risks and determining treatment plans. Considering the likelihood of cancer for hypo echoic nodules with uneven edges making precise diagnoses and implementing effective treatment approaches are essential for ensuring the best outcomes, for patients.

2.9 Breast Cancer Segmentation

It is of global concern to identify breast cancer at the earliest possible stage, as this would help increase patient survival rates [70]. Thanks to mammograms, the disease can be detected early enough for a relatively low cost. The accurate diagnosis of breast cancer is made possible through efficient segmentation in the lesion, which is vital for performing various image analysis tasks like detection, feature extraction, segmentation, and treatment planning. Once a breast image has been segmented and the tumor regions accurately delineated, the amount of tissue and the volume of the breasts can be accurately evaluated, aiding in the development of tailored treatment strategies to suit the needs of the individual patient.

It is a matter of concern globally as breast cancer kills many women and should be detected at the earliest. Through mammograms, the patient can be identified early enough, resulting in huge cost savings for patients diagnosed with cancer in its early form. The effective diagnosis of breast cancer is largely dependent on the effective segmentation which is the cornerstone of many important tasks of image analysis such as detection, feature extraction, segmentation, and treatment planning. By accurately delineating tumour areas within breast images [69], segmentation empowers healthcare providers to precisely evaluate tissue volume, enabling the development of personalized treatment plans tailored to meet the specific needs of individual patients.

Chapter 3

Proposed Methodology

Integrating advanced segmentation techniques within diagnostic systems marks a significant advancement in AI-driven disease detection. By enabling precise segmentation of anatomical structures, these systems provide critical information for accurate diagnosis and treatment planning. Through the segmentation of medical images, healthcare professionals can assess structural abnormalities and variations, which aid in identifying disease patterns and selecting optimal treatment strategies. This integration improves both the efficiency and accuracy of diagnostic workflows, leading to enhanced patient care and clinical outcomes 3.1 illustrates the overview of the proposed computer-aided diagnostic system, highlighting its segmentation capabilities and its role in supporting disease detection.

3.1 Proposed Model Architecture

In this section we present the proposed model architecture. The figure 1 shows the block diagram of the model with several key components. Starting with the Unet encoder decoder type architecture proven best for medical image classification tasks we introduce a novel parallel booster architecture along with the already prove unet encoder decoder architecture. Idea is to provide with more than one path with different kernel sizes This allows the network to adjust for the limitations of different methods of simulating a kernel. Using multiple kernel sizes helps in identifying the general region of the target while accurately detecting the edges. The model focuses on preserving boundary de-

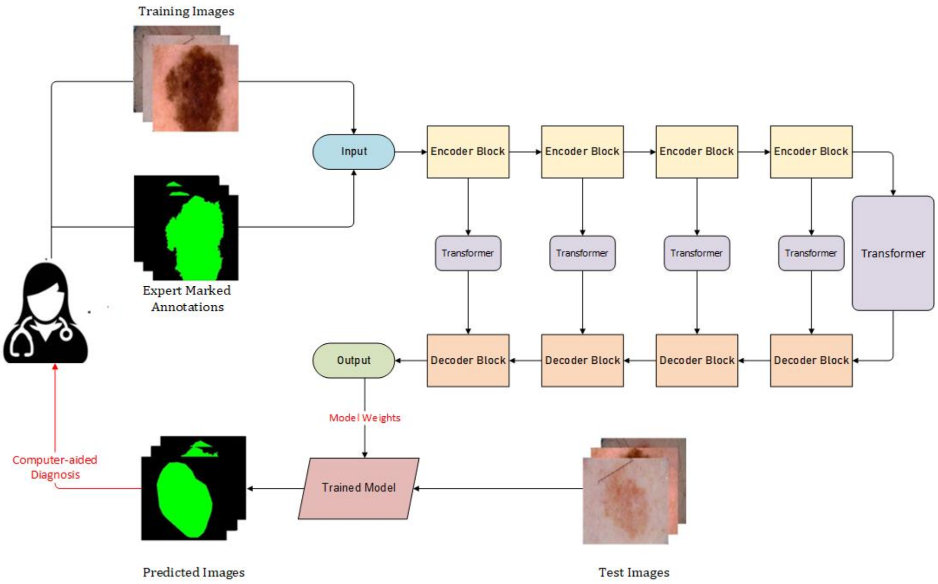

Figure 3.1: Encoder-Decoder Model for Computer-Aided Diagnosis.

tails and minimizing the use of pooling layers, as these layers often reduce the dimensions of feature maps and can lead to a loss of spatial information.

Further, we have also deployed skip connections which traditionally carry low level features such as edges texture and fine details from the encoder directly to corresponding layer of decoder and, so as to keep the complexity low and also fitted a transformer attention layer in skip connection so when it passes on information global context is also considered.

The Bottle neck enhancement is where the model captures the most abstract and global feature before moving on to the decoder stage. It acts as a bridge between encoder and decoder ensuring the model can capture deep representations of the input. This models adds TFA(transformer attention block) and ANA(attention in attention block) attention block to bottle neck. With in the (TFA) There are position encoding attention (PE) and Scaled Dot product Attention (SDP) Transformers are inherently good at capturing relationships between different parts of data but don't have a built-in sense of order or spatial structure (which is crucial for images). The position encoding is added to provide information about where each pixel is located in the input. This helps the model understand not only the features of the image but also their spatial relationships, which is essential for medical image segmentation where location is key (e.g., the boundaries of an organ). One the other hand

Scaled Dot-Product Attention mechanism works by calculating the relationship (attention score) between each pixel (or feature) and every other pixel. In segmentation, this means that the model can capture long-range dependencies, i.e., understanding how a pixel in one region of an image relates to another pixel in a distant region. This can help in identifying if the target region is present in more than one instance in the input image.

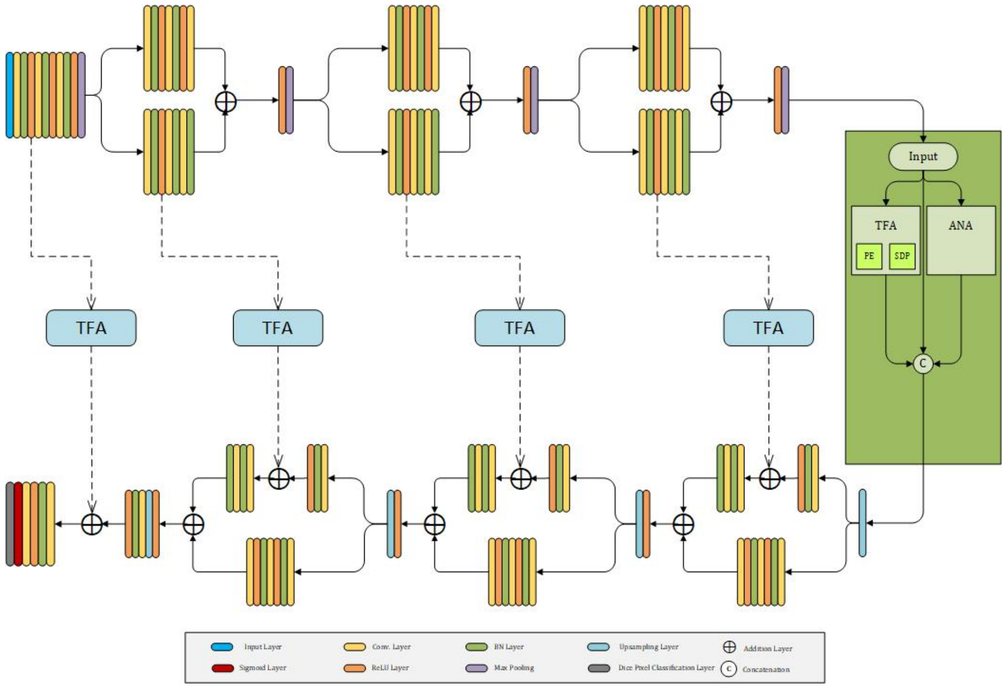

Figure 3.2: Block diagram of the proposed model.

The other important ANA Attention Block within the bottle neck builds upon the traditional attention mechanism but goes a step further by applying attention within the attention scores themselves. In simpler terms, it refines the attention process by allowing the model to focus even more selectively on important areas of the image. Finally, but not least, we have addressed the problem of class imbalance by using a Dice loss in the pixel classification layer.

In our implementation, we utilize a structure with four encoder and four decoder blocks with dense inter connections. Let $l^{n\times n}$ be the $n \times n$ convolution operation $f^{n\times n}$ following a by batch normalisation (β_n) and ReLU ($\Re$) operations for any given input ($\texttt{In}$) as defined by (Eq. 3.1.1).

$$l^{n\times n} = \Re\left(f^{n\times n}(\texttt{In})\right) \tag{3.1.1}$$

The initial skip connection (s_o) are calculated by applying the $l^{3\times 3}$ operation to the input of the network (X_{in}) as shown in (Eq. 3.3.2).

$$s_o = l^{3\times 3}(X_{in}) \tag{3.1.2}$$

Similarly, the output of the initial encoder block denoted by (E_o) is computed as (Eq. 3.3.3).

$$E_o = m_p\left(l^{3\times 3}\left(l^{3\times 3}(s_o)\right)\right) \tag{3.1.3}$$

where (m_p) defines maxpooling operation. The output of the k^{th} encoder block (E_k) is computed by (Eq. 3.3.4).

$$E_k = m_p\left[\Re\left\{\beta_n\left(f^{3\times 3}\left(\beta_n\left(f^{3\times 3}(s_k)\right)\right)\right) + f^{3\times 3}\left(l^{3\times 3}\left(l^{3\times 3}(E_{k-1})\right)\right)\right\}\right] \tag{3.1.4}$$

where (s_k) is the k^{th} skip connection and is computed as given in (Eq. 3.3.5).

$$s_k = l^{3\times 3}(E_{k-1}) \tag{3.1.5}$$

After the encoder block extracts information, two consecutive attention blocks called TFA (Transformer Attention) help in refining the extracted information. This is followed by an ANA (Attention in Attention) block, which enhances local contextual information by integrating it with global spatial information. The enhanced and refined feature information is then passed to the decoder stage, which reconstructs the spatial feature maps. Let (D_o).(Eq. 3.1.6).

$$D_o = \texttt{ANA}(E_k)©\texttt{TFA}(E_k) \tag{3.1.6}$$

Concatenation operation here is denoted by ©.We then apply TFA to skip connections and incorporate this information by performing the ($l^{3\times 3}$) operation on the input from the k^{th} decoder block, as defined in (Eq. 3.1.7)

$$\Im_k = \texttt{TFA}(s_k) + l^{3\times 3}(u_p(D_{k-1})) \tag{3.1.7}$$

Here, u_p represents the upsampling operation, which expands the spatial dimensions of the feature maps. The output of the k^{th} decoder block is computed as shown in (Eq. 3.1.8).

$$D_k = \Re\left[f^{3\times 3}\left(l^{3\times 3}\left(l^{3\times 3}\left(u_p\left(D_{k-1}\right)\right)\right)\right) + \beta_n\left(f^{3\times 3}\left(\beta_n\left(f^{3\times 3}\left(\Im_k\right)\right)\right)\right)\right] \tag{3.1.8}$$

The model's output, (X_{out}) is obtained by first applying the $l^{3\times 3}$ operation, followed by a ($f^{1\times 1}$) convolution and the sigmoid function (σ), as illustrated in (Eq. 3.1.9

$$X_{out} = \sigma(f^{1\times 1}(l^{3\times}(\Im_k))) \tag{3.1.9}$$

The final binary prediction mask is generated by applying a Dice pixel classification layer to the model's output.

3.2 TFA (Transformer Attention)

This transformer attention layer TFA function incorporates key ideas from Transformer models, particularly positional encoding and scaled dot-product attention, to enhance the network's ability to capture long-range dependencies and contextual information across spatial dimensions.

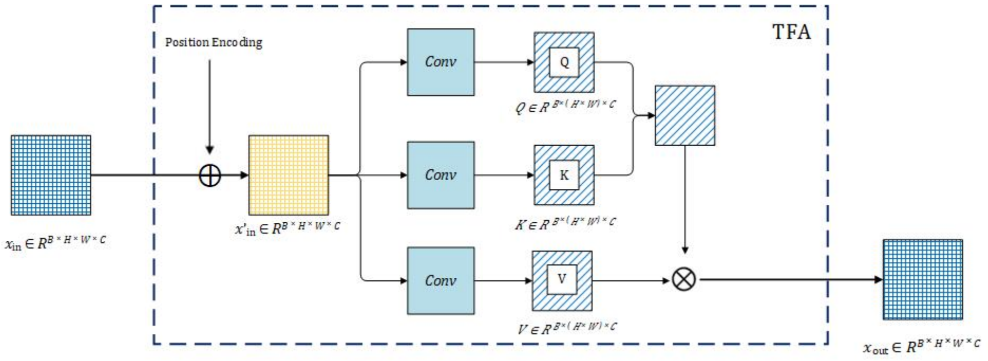
TFA
Position Encoding
Conv
Conv
Conv
Q
K
V
$Q \in R^{B\times(H\times W)\times C}$
$K \in R^{B\times(H\times W)\times C}$
$V \in R^{B\times(H\times W)\times C}$
$x_{in} \in R^{B\times H\times W\times C}$
$x'_{in} \in R^{B\times H\times W\times C}$
$x_{out} \in R^{B\times H\times W\times C}$

Figure 3.3: Schematic diagram of Transformer based Attention(TFA).

Positional Encoding: In traditional Transformer models, positional encoding is used to inject spatial information into the input, as they lack an inherent sense of position due to the self-attention mechanism. The function Position Encoding generates a matrix that encodes the position of each pixel/element in the spatial dimensions (H, W). This encoding adds information about the spatial location to the input tensor. The positional encoding (PE) is added to the input tensor, which now has both feature and spatial position information. The height dimension (H) gets a positional embedding using h embedding layer. The positions (from 0 to H-1) are fed into the height embedding layer as shown in equation 3.2.1:

$$h_{\text{embedding}} = \text{Embedding}(h_{\text{range}}), \quad h_{\text{range}} = \{0, 1, 2, \ldots, H-1\} \tag{3.2.1}$$

Similarly, the width dimension (W) gets a positional embedding using w embedding layer as shown in equation 3.2.2:

$$w_{\text{embedding}} = \text{Embedding}(w_{\text{range}}), \quad w_{\text{range}} = \{0, 1, 2, \ldots, W-1\} \tag{3.2.2}$$

The height embedding is expanded along the width dimension (W) and repeated for every width position Similarly, the width embedding is expanded along the height dimension (H) and repeated for every height position. The positional encoding from both the height and width dimensions are concatenated along the last axis.

The overall position encoding function can be written as shown in equation 3.2.3, where $E_h(i)$ is the embedding vector for position i in the height dimension, and $E_w(j)$ is the embedding vector for position j in the width dimension. Here, $i \in \{0, \ldots, H-1\}$ and $j \in \{0, \ldots, W-1\}$.

$$\text{PE}_{i,j} = [E_h(i), E_w(j)] \tag{3.2.3}$$

Further more the embeddings are learned and initialized using the HeUniform initialization method. This ensures the weights are initialized in a suitable range, improving model convergence during training.

Scaled dot product Attention: The Scaled Dot-Product Attention mechanism is an essential component in contemporary neural network architectures, such as the Transformer. This mechanism enables the network to attend to various parts of an input sequence by calculating attention scores. By using scaled dot-product attention, the transformer layer helps the model capture global context important for medical images where certain patterns or structures might span over different parts of the image. In segmentation tasks, capturing relationships between distant regions can improve the accuracy of boundary detection, reducing issues like incorrect segmentations of overlapping or complex structures.

Let: $x \in \mathbb{R}^{B\times H\times W\times C}$, where B is the batch size, H and W are the height and width of the feature map, and C is the number of channels. Then, the input feature map is reshaped into 2D spatial maps using the following equations:

- Query:

$$Q = \text{reshape}(x; B, H\times W, C) \quad \text{and then permute to} \quad Q \in \mathbb{R}^{B\times(H\times W)\times C} \tag{3.2.4}$$

- Key:

$$K = \text{reshape}(x; B, H\times W, C) \quad \text{and then permute to} \quad K \in \mathbb{R}^{B\times(H\times W)\times C} \tag{3.2.5}$$

- Value:

$$V = \text{reshape}(x; B, H\times W, C) \quad \text{and then permute to} \quad V \in \mathbb{R}^{B\times(H\times W)\times C} \tag{3.2.6}$$

The attention score is calculated as the dot product between the query $\mathbf{Q}$ and the key $\mathbf{K}$, followed by scaling by $\frac{1}{\sqrt{d_k}}$, where d_k is the temperature scaling factor, a hyperparameter controlling the scaling (which is typically the dimensionality of the key).

$$\text{energy} = \mathbf{Q}\cdot\mathbf{K}^\top \in \mathbb{R}^{B\times C\times C} \tag{3.2.7}$$

Then scaling factor is applied as shown in equation 3.2.8:

$$\text{Attention score} = \frac{\text{energy}}{\sqrt{d_k}} = \frac{\mathbf{Q}\cdot\mathbf{K}^\top}{\sqrt{d_k}} \tag{3.2.8}$$

SoftMax is applied to the attention score to convert it into probabilities as shown in equation in 3.2.9:

$$\text{Attention} = \text{SoftMax}\left(\frac{\mathbf{Q} \cdot \mathbf{K}^{\top}}{\sqrt{d_k}}\right) \in \mathbb{R}^{B \times C \times C} \tag{3.2.9}$$

Further more to prevent overfitting and regularize dropout is applied. The attention-weighted values are computed as the dot product between the value V and the normalized attention probabilities:

$$\text{output} = \mathbf{V} \cdot \text{Attention dropout} \in \mathbb{R}^{B \times (H \times W) \times C} \tag{3.2.10}$$

Finally, the output is reshaped back to match the original spatial dimensions, which is the same as the input shape $\mathbb{R}^{B \times (H \times W) \times C}$.

3.3 ANA (Attention-in-Attention) Attention Block

This custom ANA attention block is used for enhancing feature representations. The purpose of this block is to leverage attention mechanisms to enhance key features in the input while reducing irrelevant or redundant information. ANA helps the model not only focus on relevant areas but also refine its understanding of the **semantic importance** of different features. This is especially useful in medical images where small, intricate features need to be captured, and where misclassification could lead to incorrect diagnosis. Instead of a single pass of attention computation, ANA applies nested attention mechanisms, where the model first learns a coarse attention map and then refines it further with another layer of attention. This means that the model can gradually refine its focus on key regions of the image, particularly useful for detecting fine-grained details in medical images like small tumors or subtle changes in tissue.

The process starts with an input tensor I with dimensions $H \times W \times C$, where H, W, and C are the height, width, and channels of the input on which operations are applied. Let BN(.) denote batch normalization, $\sigma(.)$ denote ReLU activation, and Conv2D(.) represent 2D convolution. The output of the convolution and normalization is the feature map F_1 as shown in equation 3.3.1, where K is the number of filters per class and n_{classes} is the number of output classes.

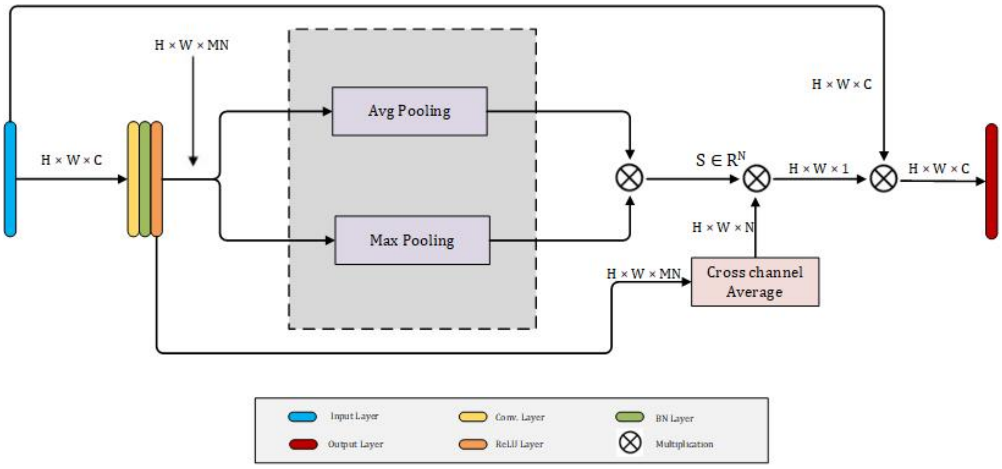

Figure 3.4: Schematic diagram of Attention in Attention(ANA).

$$F_1 = \sigma\left(\mathrm{BN}\left(\mathrm{Conv2D}\left(I, k \cdot n_{\mathrm{classes}}, 3 \times 3\right)\right)\right) \tag{3.3.1}$$

Then the pooling operation generates two summary statistics from the feature map $F_2 = F_1$, which are $x_1 = \mathrm{GlobalMaxPooling}(F_2)$ and $x_2 = \mathrm{GlobalAveragePooling}(F_2)$. $\mathrm{GlobalMaxPooling}(F_2)$ takes the maximum value across spatial dimensions for each channel, whereas $\mathrm{GlobalAveragePooling}(F_2)$ takes the mean value across spatial dimensions for each channel. Then, element-wise multiplication of the pooling outputs is performed; this way, the pooled outputs are combined as shown in equation 3.3.2.

$$x = x_1 \odot x_2 \tag{3.3.2}$$

Further, the pooled feature vector x is reshaped and averaged along the last dimension to obtain the attention map S. In equation $S = \mu(x, \mathrm{axis} = -1)$, μ denotes the mean operation along the last dimension (which corresponds to k, the number of filters per class). This results in the attention map S, which indicates the importance of each class across the spatial dimensions.

The feature maps $F1$ are reshaped to apply the attention map as shown in equation 3.3.3

$$x = \mathrm{Reshape}(F_1, (H, W, n_{\mathrm{classes}}, k)) \tag{3.3.3}$$

Afterward, the mean is computed along the last dimension again to collapse the channel dimension: $x = \mu(x, \text{axis} = -1)$. This operation reduces the feature maps from a higher-dimensional space to the attention space. The attention map S is applied to the feature maps by element wise multiplication as shown in equation 3.3.4:

$$x = S \odot x \tag{3.3.4}$$

The resulting map x is averaged along the last dimension to obtain the final attention map M. The final output is obtained by element-wise multiplication between the input I and the attention map M. as shown in equation 3.3.5.

$$\text{semantic} = I \odot M \tag{3.3.5}$$

Chapter 4

Experimental Setup

In this section, we first we provide a thorough overview of the datasets used in this thesis research on medical image segmentation, followed by an in-depth examination of the experimental methodology applied to assess the proposed algorithm.

4.1 Datasets

The efficacy of the proposed model was evaluated using a diverse set of publicly accessible datasets, consisting of both dermoscopic and ultrasound images. Four benchmark datasets were utilized, with two datasets containing skin images and two containing ultrasound images. The ISIC challenge includes three primary tasks which are localization and detection of visual features/patterns, lesion segmentation and disease classification. Although these tasks collectively aim to automate melanoma diagnosis from dermoscopic images, the primary focus of this research lies in the segmentation task. Lesion segmentation is essential for precisely outlining the boundaries of skin lesions, forming the foundation for further analysis.

Description in a detaile of each dataset is provided below, with their distribution summarized in Table 4.1. This table offers information about the datasets used to assess the performance of proposed model, including the number of images in the training, validation, and test sets, as well as image resolution. The "Segmentation Masks" column highlights whether segmentation masks are available

Datasets	Number of Images			Resolution	Segmentation Masks
	Train	Validation	Test		
BUSI [69]	780	N.A	80	500×500	Available
DDTI [71]	637	N.A	N.A	245×360 - 560×360	Available
ISIC2017 [73]	2000	N.A	600	679×453 - 6748×4499	Available
ISIC2016 [72]	900	N.A	379	679×453 - 6748×4499	Available
PH 2 [75]	200	N.A	N.A	768×560	Available

Table 4.1. Medical Image Segmentation Dataset's Description for the proposed model's evaluation.

for each dataset. Only datasets that included segmentation masks were used for segmentation tasks in this evaluation. The BUSI [69], DDTI [71], ISIC2016 [72], ISIC2017 [73] ISIC2018 [74] and PH2 [75] datasets all contain segmentation masks, making them suitable for segmentation evaluations.

In contrast, the ISIC2019 [76] and ISIC2020 [77] datasets, commonly utilized in dermatology for skin lesion analysis, do not provide segmentation masks for the lesions. Segmentation masks are crucial for tasks like boundary delineation and pixel-level classification. As a result, these datasets were excluded from the segmentation evaluation and only used for detection tasks. Despite the absence of segmentation masks [78], the ISIC2019 and ISIC2020 datasets remain valuable for other tasks, such as lesion classification, where the goal is to differentiate between benign and malignant lesions based on their visual and clinical characteristics. These datasets contribute to the broader objective of enhancing the diagnosis and treatment of skin lesions through computer-aided analysis. However, challenges such as class imbalance in multi-detection and image duplication issues in the ISIC2020 dataset 78, resulting from the merging of multiple datasets for detection tasks, were also noted

BUSI: The BUSI dataset [69] contains 780 images of breast ultrasound collected from women aged 25 to 75. These images are provided in PNG format with an average resolution of 500×500 pixels, and for uniformity, images are of size 256×256 pixels. This dataset includes ground truth annotations across three categories: normal, benign, and malignant. A three-fold cross-validation approach is implemented to assess the performance of segmentation algorithms.

Breast cancer is a leading cause of death among women globally, highlighting the critical need for early detection to improve survival rates. The BUSI dataset provides essential data for the diagnosis of breast cancer through ultrasound imaging. With its categorization into normal, benign, and malignant classes, the dataset supports various tasks such as classification, detection, and segmentation of breast cancer. By utilizing this dataset, researchers can develop and refine machine learning algorithms to enhance breast cancer diagnosis and ultimately improve patient outcomes.

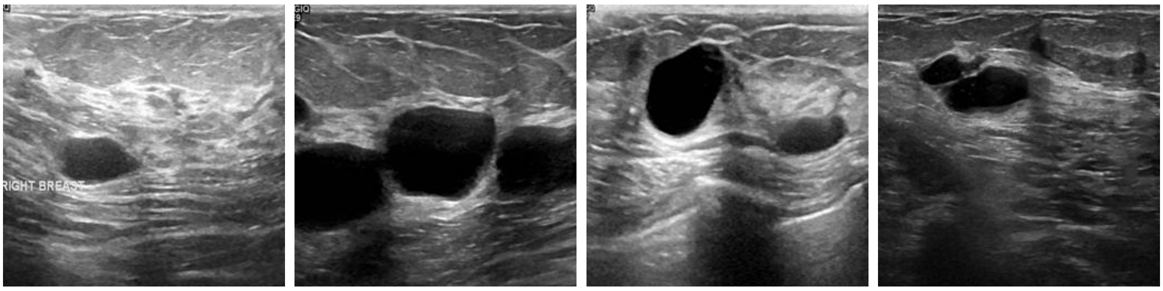

Figure 4.1: BUSI Dataset images.

DDTI: The DDTI dataset [71] includes 637 ultrasound images of thyroid nodules in PNG format. To maintain consistency in image size, images in the dataset have been resized to 256×256 pixels. The dataset is organized into training, validation, and test sets, using an 80%, 10%, and 10% split. Additionally, a three-fold cross-validation approach is applied to ensure a robust and reliable evaluation of the segmentation algorithm.

This online collection of thyroid ultrasound pictures is provided by the collaboration, between the National University of Colombia and IDIME (Medical Diagnostic Institute Colombia) offering an asset for researchers in the field of science and medicine to utilize for work on computer assisted detection (CAD) systems designed for assessing thyroid nodules efficiently and effectively in medical imaging analysis tasks like this. Not only does it support algorithm development but also serves as a valuable resource for educating and training new radiologists with a wide range of thyroid ultrasound images, at their disposal. By ensuring that this information is readily available, to all parties the dataset fosters the advancement of studies and cooperation, in scrutinizing thyroid nodules which in turn enhances the precision of diagnoses and the quality of healthcare provided to patients.

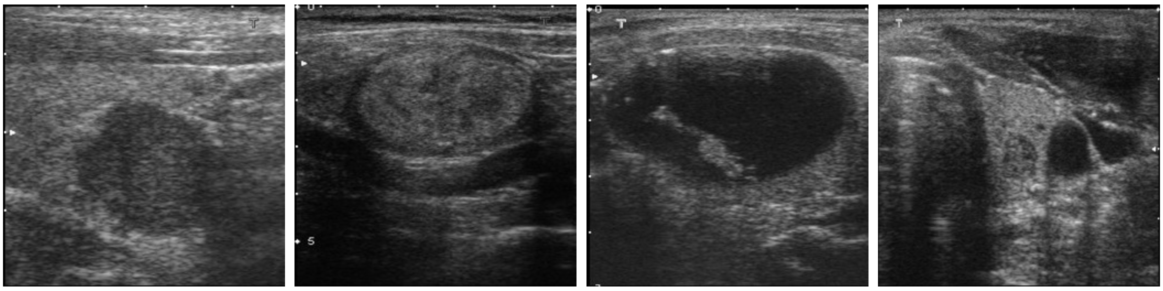

Figure 4.2: DDTI Dataset images.

ISIC 2016: ISIC 2016 dataset [72] consists of 900 dermoscopic images allocated for training, along with 379 images reserved for testing. Each image is accompanied by ground truth masks, which provide critical annotations for assessing segmentation performance. Dermoscopic images present significant challenges due to the inherent complexity in their structure. These images often display irregular boundaries, diverse shapes, contrasting textures, and may include scars or artifacts, all of which make accurate segmentation difficult. Additionally, the dataset include a variety of skin lesions, such as melanoma, nevi, and other benign or malignant conditions, each with distinct features that add to the segmentation complexity.

Addressing these challenges requires sophisticated segmentation algorithms capable of accurately identifying and delineating lesion boundaries, despite irregularities and the presence of artifacts. The varying shapes, textures, and boundaries, combined with factors like inconsistent lighting conditions, image quality differences, and patient demographic diversity, heighten the complexity of the dataset. Advanced segmentation techniques are essential for overcoming these obstacles and producing reliable results.

Furthermore, rigorous evaluation and validation methodologies are necessary to effectively gauge the performance of segmentation algorithms in handling such complex data. The robustness of segmentation models must be assessed thoroughly to ensure their applicability in real-world clinical environments. The ISIC 2016 dataset, with its comprehensive annotations and challenging image characteristics, provides an ideal platform for training and testing medical image segmentation models, ensuring they perform effectively in clinical practice.

ISIC 2017: The ISIC 2017 dataset [73] comprises 2000 dermoscopic images, providing a substan-

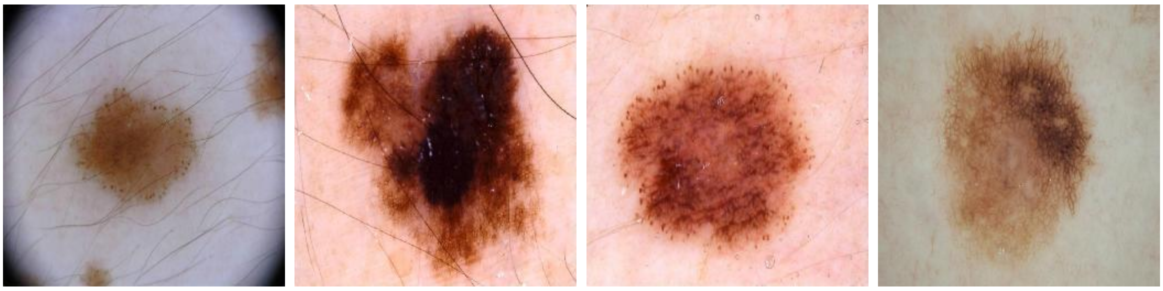

Figure 4.3: ISIC 2016 Dataset images.

tial amount of data for the development and training of segmentation models. In addition, it includes 150 images reserved for validation and 600 images specifically allocated for testing and performance evaluation. This dataset serves as an important resource for both researchers and clinicians in dermatology, offering a diverse and well-annotated collection of images for advancing the development of medical image segmentation frameworks.

One of the key challenges in skin lesion segmentation lies within the ISIC 2017 Skin Lesion dataset, particularly in its focus on melanoma detection. This dataset is considered one of the most difficult among the ISIC challenges, including ISIC 2016 and ISIC 2018, due to the complexity and variability of the lesions present in the images. The ISIC 2017 challenge has become a benchmark for evaluating the effectiveness of segmentation models on real-world dermoscopic images, pushing the limits of algorithm development and validation.

The diverse range of lesions, including those with irregular shapes, varying sizes, and textures, makes it a valuable tool for testing segmentation algorithms in real-world clinical conditions. Researchers commonly use this dataset to assess the performance of their models, taking advantage of the extensive annotations and variety of lesion types[79]. By providing such a challenging dataset, the ISIC 2017 challenge plays a crucial role in fostering the development of segmentation algorithms that can accurately delineate lesion boundaries, thereby improving diagnostic accuracy and patient outcomes in clinical settings.

PH2: The PH2 dataset [75] consists of a carefully curated set of 200 dermoscopic images, each paired with corresponding ground truth masks. This dataset provides an important resource for developing and assessing skin lesion segmentation models. Its diverse representation of lesion types,

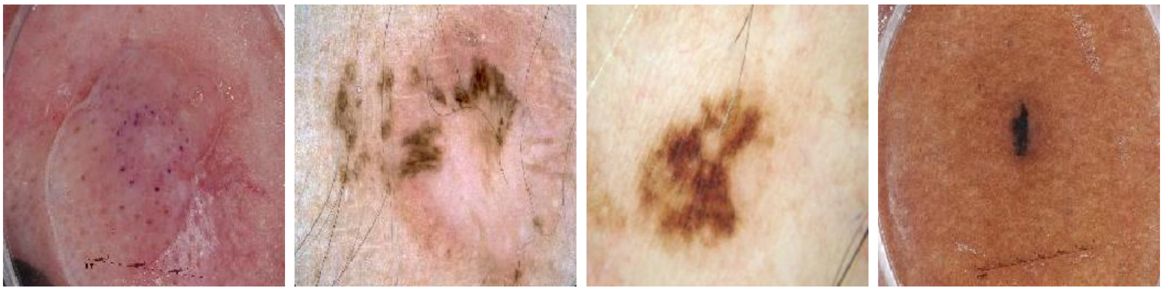

Figure 4.4: ISIC 2017 Dataset images.

including various benign and malignant conditions, enables comprehensive analysis. The inclusion of high-quality annotations ensures that models trained on this dataset can be rigorously evaluated for accuracy in boundary delineation and pixel-wise classification, making it highly valuable for research in dermatological image analysis.

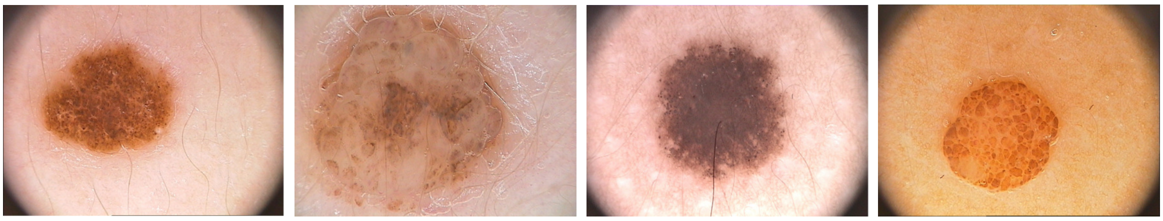

Figure 4.5: PH2 Dataset images.

4.2 Model Bench-marking

Benchmarking of medical image segmentation models has a pivotal role in evaluating their performance and guiding future research efforts. In this study, we adhered to well-established methodologies presented by frameworks like Ms Red [80], FTN Network [81], and DconnNet [82], which provide a standardized approach for assessing segmentation algorithms. These guidelines ensure uniformity, enabling fair comparisons across different studies and models.

For our benchmarking analysis, we thoughtfully chose a varied collection of state-of-the-art models known for their outstanding effectiveness in medical image segmentation. We structured our training setup following the approaches outlined in works such as Ms Red [80] and ARU-GD [83]. The models chosen for evaluation include Swin-Unet [84], U-Net [52], ARU-GD [83], Att-UNet [85], UNet++ [86], DuckNet [**105**], and Meta-Poly [87], selected for their innovative architectures and proven results in previous research.

Once these models were trained on relevant datasets, we rigorously evaluated their segmentation outputs. Our analysis focused on key performance metrics like accuracy, sensitivity, specificity, and the Dice similarity coefficient [80], providing a detailed view of each model's strengths and areas for improvement. These insights enabled us to make well-informed decisions regarding which models are best suited for particular clinical applications.

To ensure a thorough comparison, we also reviewed and integrated benchmarking data from previous studies on skin lesion segmentation. Models like Ms Red [80], FAT-Net [88], and AS Net [89] have been tested on widely-used datasets such as ISIC 2016, ISIC 2017, ISIC 2018, and PH2. These datasets, when used as out-of-distribution benchmarks, helped us evaluate the generalization capability and robustness of the models across varying data distributions.

By comparing our selected models against these established results, we gained a deeper understanding of the models' relative strengths in different scenarios. This comprehensive benchmarking contributes to the broader domain of medical image analysis and AI-driven diagnostic tools

4.3 Performance measure

The evaluation of our model's performance utilizes five metrics: Accuracy, Specificity, Jaccard index (IOU), Dice coefficient and Sensitivity . These metrics provide a well-rounded evaluation of the model's ability to accurately segment skin lesions. This allows a comprehensive assessment of the model's accuracy and effectiveness in various segmentation scenarios.

Jaccard Index (IoU)

The Jaccard index, also known as Intersection over Union (IoU), quantifies the overlap between two sets by dividing the size of their intersection by the size of their union. This metric is computed using 4.3.1, where true positives (TP) indicate correctly identified pixels, false positives (FP) represent incorrectly identified pixels, and false negatives (FN) are missed pixels. The Jaccard index provides a reliable measure of similarity between predicted and actual segmentation outputs.

$$\text{Jaccard Index (IoU)} = \frac{TP}{TP+FP+FN} \tag{4.3.1}$$

Dice Coefficient (Ds)

The Dice coefficient (Ds) is a commonly used metric for assessing similarity in image segmentation tasks. It quantifies the agreement between the predicted segmentation mask and the ground truth by calculating twice the size of their intersection, then dividing that by the total number of elements in both sets combined. The formula, given in 4.3.2, incorporates true positives (TP), false positives (FP), and false negatives (FN) to assess how well the segmentation aligns with the actual results. This metric is particularly effective in highlighting the overlap between the two sets.

$$\text{Dice Coefficient (Ds)} = \frac{2 \times TP}{2 \times TP+FP+FN} \tag{4.3.2}$$

Accuracy (Acc)

Accuracy (Acc) reflects the percentage of correctly identified instances, including both positive and negative cases, from the overall dataset. It is computed using Equation 4.3.3, with true positives (TP) representing correctly classified positive samples, true negatives (TN) for correctly classified negative samples, and false positives (FP) and false negatives (FN) indicating misclassifications. This metric provides an overall measure of the model's ability to make accurate predictions across all classes.

$$\text{Accuracy (Acc)} = \frac{TP+TN}{TP+TN+FP+FN} \tag{4.3.3}$$

Sensitivity (Sn)

Sensitivity (Sn), also known as recall or the true positive rate, measures the model's effectiveness in accurately identifying positive instances among the total number of actual positive cases. It is calculated using 4.3.4, where true positives (TP) refer to correctly identified positive cases, and false negatives (FN) indicate positive instances that were missed. This metric assesses how effectively the model captures all relevant positive samples

$$\text{Sensitivity (Sn)} = \frac{TP}{TP+FN} \tag{4.3.4}$$

Specificity (Sp)

Specificity (Sp) quantifies the proportion of true negative samples accurately identified as negative from the total number of actual negative samples. It is determined using Equation 4.3.5, where TN denotes true negatives and FP signifies false positives.

$$\text{Specificity (Sp)} = \frac{TN}{TN+FP} \tag{4.3.5}$$

4.4 Implementation Details

The preprocessing phase begins by resizing all training images to a uniform dimension of 256×256 before feeding them into the model. For optimization, the Adam optimizer is applied with parameters $\beta_1 = 0.90$ and $\beta_2 = 0.999$. These values are chosen based on previous research findings, as highlighted in [90], where these specific parameters have proven effective in medical image segmentation tasks.

To prevent overfitting, an Early Stopping mechanism is introduced, starting from the 10^{th} epoch. Notably, the proposed loss function is responsive to dynamic weighting, which helps the model minimize the loss more effectively. However, in some cases, this can negatively impact the Jaccard index. To improve performance across diverse datasets, the Jaccard coefficient is manually selected as the monitoring metric when the L_{bl} label is in use. Otherwise, the model monitors the validation loss.

All experiments are run using Google Colab Pro, leveraging a T4 GPU with a batch size 12. The implementation is carried out in Keras framework, using Python 3.10.12.

Chapter 5

Result and Discussion

5.1 Ablation Study of Proposed Model

The comprehensive evaluation of the proposed model was done using ISIC 2017 dataset, with the quantitative results highlighting performance improvements presented in Table 5.1. To evaluate the impact of various components on enhancing the baseline UNet-based CNN model, a carefully designed ablation study was carried out. Initially, the baseline model was used as a reference, and the results were computed based on its performance. In the second phase of experimentation, the TFA (Transformer Attention) module was incorporated into both the skip connections and the bottleneck. In the third phase, the novel ANA (Attention IN Attention) block was added to the bottleneck in conjunction with the TFA module. This combination demonstrated a significant boost in overall performance. The synergistic effect of these components proved to be especially effective in enhancing the model's capabilities.

This section further presents a performance comparison of the proposed model against recent methods across various datasets, including BUSI[69], DDTI[71], ISIC 2016 [72], ISIC 2017 [73], and PH2 [75]. Comparisons in the tables are derived from the cited literature. These findings demonstrate the generalization capability of the proposed model across diverse datasets, particularly for ultrasound images: the BUSI dataset for breast cancer segmentation and the DDTI dataset for thyroid nodule segmentation. This generalization underscores the model's adaptability to other medical

Method	Performance Measures (%)				
	$J_{ind} \uparrow$	$D_s \uparrow$	$Acc \uparrow$	$S_n \uparrow$	$S_p \uparrow$
Base Network (BN)	79.74	86.72	94.50	87.17	92.33
BN + TFA	81.76	88.60	95.54	88.65	94.81
BN + TFA + ANA	83.88	89.89	95.95	89.85	95.37

Table 5.1. Ablation study of proposed model on ISIC 2017 dataset.

image segmentation modalities

5.2 Performance Comparison with other models on the BUSI dataset

To evaluate the performance of our model for breast cancer segmenttion A publically avalaible BUSI dataset [69] is used.Comparisons are made against several state of the art models, such as U-Net [52], FPN [90], Swin-Unet [84] etc. Table 5.2 provides a detailed statistical comparison between the proposed model and these methods. The proposed model demonstrates improvement in the Jaccard index, achieving a higher score on the BUSI dataset [69] compared to other approaches. Additionally, this model is tested on images related to breast cancer which presents challenges such as varying sizes and irregular shapes.

5.3 Performance Comparisons with other models on the DDTI dataset

Performance of the model for thyroid nodule segmentation is evaluated using the publicly available DDTI dataset [71]. Comparison is done with state of the art methods, including U-Net [52], Attention U-Net [85], Swin-Unet [84] etc.Table 5.3 presents a statistical comparison of proposed model with these methods. Results in the table show that proposed model achieves a significant improvement in the parameter Jaccard index on the DDTI dataset. The proposed model is also evaluated on thyroid nodule images that present challenges like irregular shapes and varying sizes.

Methods	**Performance Measures (%)**				
	$J_{ind} \uparrow$	$D_s \uparrow$	$Acc \uparrow$	$S_n \uparrow$	$S_p \uparrow$
ConvEDNet[91]	73.57	82.70	-	85.51	-
U-Net[52]	67.77	76.96	95.48	78.33	96.13
DeeplabV3+[92]	73.48	82.68	-	83.37	-
BCDU-Net[57]	74.49	66.75	94.82	86.85	95.57
UNet++[53]	76.85	76.22	97.97	78.61	98.86
BGM-Net[93]	75.97	83.97	-	83.45	-
Swin-Unet[84]	77.16	84.45	97.55	84.81	98.34
Meta-Poly[87]	77.93	-	97.81	89.17	-
DuckNet[86]	85.63	-	98.69	**95.36**	-
ARU-GD[83]	77.07	83.64	97.94	83.80	98.78
Our model	**87.97**	**93.36**	**99.05**	93.86	**99.41**

Table 5.2. Performance comparison of Our model on (BUSI) breast ultrasound images dataset.

Methods	**Performance Measures (%)**				
	$J_{ind} \uparrow$	$D_s \uparrow$	$Acc \uparrow$	$S_n \uparrow$	$S_p \uparrow$
BCDU-Net[57]	57.79	69.49	93.22	78.31	94.34
DuckNet[86]	68.32	-	94.17	77.47	-
MShNet[74]	73.43	75.01	-	82.21	-
U-Net[52]	74.76	84.08	96.55	85.50	97.57
UNet++[53]	74.76	84.08	96.55	85.50	97.57
Meta-Poly[87]	76.64	-	96.35	89.31	-
Swin U Net[84]	75.45	84.87	96.93	86.42	97.98
Attention U-Net[85]	77.37	84.91	-	81.70	-
M-Net[94]	79.38	86.40	-	75.45	-
nnUnet[2]	80.76	88.59	-	85.23	-
ARU-GD[83]	77.07	83.64	97.94	83.80	98.78
DeeplabV3+[92]	82.66	87.72	-	79.54	-
N-Net[94]	88.46	92.67	-	91.94	-
Our model	**93.09**	**96.18**	**99.16**	**96.28**	**99.46**

Table 5.3. Performance Comparison of Our model on DDTI Thyroid Nodule Segmentation Dataset.

5.4 Performance Comparison on the ISIC 2016 dataset

A comprehensive analysis is done when evaluating proposed model on ISIC 2016 dataset with other state of the art models. To ensure fairness, all models were tested in identical computing environments with consistent data augmentations. The methods included for comparison were Ms RED [80], Swin-Unet [84], UNet++ [53], CPFNet [58], BCDU-Net [57], DAGAN [95], FAT-Net [88], ARU-GD [83], U-Net [52], and Hyper-Fusion Net [96]. It's important to note that, aside from Swin-Unet, UNet++, BCDU-Net, U-Net, and ARU-GD, the results for other models were taken from their respective cited papers. The proposed model demonstrated superior performance, with a Jaccard index improvement ranging from 0.93% to 7.72% compared to other methods. As shown in Table 5.4, proposed model consistently surpassed the competing models across all evaluation metrics. These comparisons highlight proposed models superior performance, particularly in challenging cases involving lesions on skin of varying sizes and irregular shapes, where it consistently delivers the best segmentation outcomes.

Method	**Performance Measures in (%)**				
	$J_{ind} \uparrow$	$D_s \uparrow$	$Acc \uparrow$	$S_n \uparrow$	$S_p \uparrow$
BCDU-Net[57]	83.43	80.95	91.78	78.11	96.20
ARU-GD[83]	85.12	90.83	94.38	89.86	94.65
CPFNet[58]	83.81	90.23	95.09	92.11	95.91
U-Net[52]	81.38	88.24	93.31	87.28	92.88
Ms RED[80]	87.03	92.66	96.42	-	-
UNet++[53]	82.81	89.19	93.88	88.78	93.52
DAGAN[97]	84.42	90.85	95.82	92.28	95.68
DuckNet[86]	85.55	-	95.64	91.40	-
FAT-Net[88]	85.30	91.59	96.04	92.59	96.02
Meta-Poly[87]	85.60	-	96.08	91.72	-
Hyper-Fusion Net[96]	88.17	-	96.64	94.22	96.45
Swin-Unet[84] 84	87.60	88.94	96.00	92.27	95.79
Our model	**90.63**	**94.48**	**97.45**	**95.49**	**96.82**

Table 5.4. Performance Comparison of proposed model on ISIC 2016 Dataset.

5.5 Performance Comparisons on the ISIC 2017 dataset

A comprehensive analysis is done when evaluating proposed model on ISIC 2017 dataset with other state of the art models. All assessments were performed under identical computing conditions with consistent data augmentation to ensure a fair comparison. The methods included for evaluation were FAT-Net [88], DAGAN [95], Ms RED [80], U-Net [52], Swin-Unet [84], UNet++ [53], BCDU-

Net [57], AS-Net [89], SEACU-Net [98], ARU-GD [83], and BA-Net [99]. It is important to note that, aside from Swin-Unet, UNet++, BCDU-Net, U-Net, and ARU-GD, the performance results for the remaining models were sourced from their respective published papers. As shown in Table 5.5, Proposed model outperformed the competing models across most evaluation metrics. The results consistently highlighted proposed model superior performance, particularly in challenging cases involving lesions over skin of varying sizes and irregular shapes, where the model closely matched the ground truth in its segmentation outcomes.

Methods	**Performance Measures (%)**				
	$J_{ind}\uparrow$	$D_s\uparrow$	$Acc\uparrow$	$S_n\uparrow$	$S_p\uparrow$
AS-Net[89]	80.51	88.07	94.66	89.92	95.72
DAGAN[95]	75.94	84.25	93.26	83.63	97.24
U-Net[52]	75.69	84.12	93.29	84.30	93.41
Meta-Poly[87]	78.31	-	94.15	85.53	-
UNet++[53]	78.58	86.35	93.73	87.13	94.41
FAT-Net[88]	76.53	85.00	93.26	83.92	**97.25**
Hyper-Fusion Net[96]	83.70	-	95.80	**92.33**	96.16
SEACU-Net[98]	80.50	89.11	95.35	-	-
ARU-GD[83]	80.77	87.89	93.88	88.31	96.31
Ms RED[80]	78.55	86.48	94.10	-	-
Swin-Unet[84]	80.89	81.99	94.76	88.06	96.05
BCDU-Net[57]	79.20	78.11	91.63	76.46	97.09
BA-Net	81.00	88.10	94.60	89.70	96.60
DuckNet[86]	82.08	-	95.41	89.03	-
Our model	**83.88**	**89.89**	**95.95**	89.85	95.37

Table 5.5. Performance Comparison of proposed model on ISIC 2017 Dataset..

5.6 Performance Comparison with other models on the PH2 dataset

At last, generalization capability of our model is evaluated through cross dataset validation. The model is trained on the ISIC 2016 dataset and tested on the PH2 dataset [75]. The performance of the proposed model on the PH2 dataset is compared against several state of the art methods, such as ICLNet [98], DCL-PSI [100], MFCN [101], and AS-Net [89]. Table 5.6 illustrates the performance comparison between the proposed model and these advanced methods. IN comparison with state of the art techniques, the Jaccard index for the proposed model shows an improvement of 3.88% to 7.79% on the PH2 dataset [75]

Method	Performance Measures in (%)				
	$J_{ind} \uparrow$	$D_s \uparrow$	$Acc \uparrow$	$S_n \uparrow$	$S_p \uparrow$
AS-Net[89]	87.60	93.05	95.20	**96.24**	94.31
ICL-Net[99]	87.25	92.80	96.32	95.46	**97.36**
MFCN[101]	83.99	90.66	94.24	94.89	93.98
DCL-PSI[100]	85.90	92.10	95.30	96.23	94.52
Our model	**90.08**	**94.16**	**96.65**	95.51	93.67

Table 5.6. Performance Comparison of Our model on PH2 Dataset.

5.7 Computational Complexity Analysis

With in this subsection, we perform an in-depth evaluation of the computational demands of the proposed model, comparing its requirements to those of current state of the art approaches. The computational comparison presented in Table 5.7, emphasizes effectiveness of the proposed approach.

The proposed model exhibits exceptional computational efficiency, notably due to its considerably lower number of learnable parameters. With just 0.68 million parameters, it surpasses other algorithms regarding parameter efficiency. Crucially, this optimization does not compromise the model's high performance in medical imaging analysis. It achieves an optimal balance between computational efficiency and excellent segmentation results. Additionally, it requires just 2.6 billion floating-

Method	**Computational Complexity Analysis**		
	Param (M) ↓	**FLOPs (G) ↓**	**Inference Time (ms) ↓**
BCDU-Net [57]	28.8	38.22	28.07
UNet++ [53]	34.9	35.6	31.3
ARU-GD [83]	33.3	33.93	29.49
U-Net [52]	32.9	33.39	28.87
DeepLabv3 [90]	37.9	33.89	29.62
Swin U-Net [84]	29	25.4	25.6
Our model	**0.68**	**2.6**	**15.4**

Table 5.7. Computational Complexity Analysis of proposed model on a spatial dimension of image 256×256.

point operations (FLOPs) and achieves an inference time of only 15.4 milliseconds. This compact design enhances the practicality of deploying the LSSF-Net method in real clinical environments. The model's smaller size makes it both efficient and effective for medical imaging tasks, facilitating easier integration and real-time use.

Chapter 6

Conclusion

In conclusion, this thesis research marks a significant leap forward in the domain of medical image segmentation, presenting effectiveness and efficacy of the proposed architecture. Keeping in view the limitations of medical image segmentation we develop a hybrid model in order to keep focus on both global and local features, a CNN based parallel booster encoder decoder to capture long range dependencies along with feature extraction and novel attention mechanisms at bottleneck and skip connections in order for extraction of global features which ultimately helped in feature enhancement and improved output results. Addition of ANA attention which focus on selectively more important features and TFA attention whose one of the key purpose is to find patterns those might span over different parts of an image helped us in achieving excellent results on benchmark datasets, confirming that the proposed method outperforms existing segmentation approaches.

To assess the versatility and generalization of proposed model, performance comparisons are done across a range of skin types and lesion characteristics, as well as in breast cancer segmentation and thyroid nodule segmentation. The results in section 5 showed consistent performance of proposed model across different types of datasets. These findings contribute significantly to ongoing efforts aimed at enhancing the precision and efficiency of diagnostic tools in medical image segmentation. Looking forward, there exists ample opportunity for further exploration and refinement of the proposed model architecture.

In summary, the advancements achieved through this research highlight the promising prospects of our model in pushing the boundaries of medical image segmentation, with far-reaching advantages in enhancing diagnostic accuracy and improving patient care in health care industry.

Chapter 7

Recommendations

Looking ahead, we advocate for ongoing exploration and refinement of the proposed model architecture to tackle the evolving challenges in medical image segmentation. This could include investigating innovative techniques for feature extraction and integration, as well as examining alternative network architectures to further boost the model's performance. Additionally, incorporating multimodal data sources, such as clinical metadata or histopathological images, may provide valuable supplementary information to enhance segmentation accuracy and reliability.

Moreover, collaboration with dermatologists and other healthcare professionals is crucial to validate the practical applicability of the proposed model in clinical environments. Conducting thorough validation studies on diverse patient populations and real-world datasets will ensure that the model meets rigorous standards for diagnostic accuracy and reliability. Fostering interdisciplinary partnerships among researchers, clinicians, and industry stakeholders will facilitate the integration of the proposed model into routine clinical practice. Ultimately, this collaboration will benefit patients by improving diagnostic outcomes and supporting more effective treatment strategies.

Bibliography

[1] Miguel Monteiro et al. “Multiclass semantic segmentation and quantification of traumatic brain injury lesions on head CT using deep learning: an algorithm development and multi-centre validation study”. In: *The Lancet Digital Health* 2.6 (2020), e314–e322.

[2] Fabian Isensee et al. “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation”. In: *Nature methods* 18.2 (2021), pp. 203–211.

[3] Fabian Isensee and Klaus H Maier-Hein. “An attempt at beating the 3D U-Net”. In: *arXiv preprint arXiv:1908.02182* (2019).

[4] Qiangguo Jin et al. “RA-UNet: A hybrid deep attention-aware network to extract liver and tumor in CT scans”. In: *Frontiers in Bioengineering and Biotechnology* 8 (2020), p. 605132.

[5] Spyridon Bakas et al. “Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge”. In: *arXiv preprint arXiv:1811.02629* (2018).

[6] Amber L Simpson et al. “A large annotated medical image dataset for the development and evaluation of segmentation algorithms”. In: *arXiv preprint arXiv:1902.09063* (2019).

[7] Nicholas Heller et al. “The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes”. In: *arXiv preprint arXiv:1904.00445* (2019).

[8] Konstantinos Kamnitsas et al. “Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation”. In: *Medical image analysis* 36 (2017), pp. 61–78.

[9] Ze Liu et al. “Swin transformer: Hierarchical vision transformer using shifted windows”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 10012–10022.

[10] Liang-Chieh Chen et al. “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs”. In: *IEEE transactions on pattern analysis and machine intelligence* 40.4 (2017), pp. 834–848.

[11] Wenqi Li et al. “On the compactness, efficiency, and representation of 3D convolutional networks: brain parcellation as a pretext task”. In: *Information Processing in Medical Imaging: 25th International Conference, IPMI 2017, Boone, NC, USA, June 25-30, 2017, Proceedings 25*. Springer. 2017, pp. 348–360.

[12] Yingda Xia et al. “3d semi-supervised learning with uncertainty-aware multi-view co-training”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2020, pp. 3646–3655.

[13] Jun Fu et al. “Dual attention network for scene segmentation”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 3146–3154.

[14] Jonathan Long, Evan Shelhamer, and Trevor Darrell. “Fully convolutional networks for semantic segmentation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3431–3440.

[15] Olaf Ronneberger, Philipp Fischer, and Thomas 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*. Springer. 2015, pp. 234–241.

[16] Md Zahangir Alom et al. “Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation”. In: *arXiv preprint arXiv:1802.06955* (2018).

[17] Ozan Oktay et al. “Attention u-net: Learning where to look for the pancreas”. In: *arXiv preprint arXiv:1804.03999* (2018).

[18] Xiao Xiao et al. “Weighted res-unet for high-quality retina vessel segmentation”. In: *2018 9th international conference on information technology in medicine and education (ITME)*. IEEE. 2018, pp. 327–331.

[19] Md Zahangir Alom et al. “Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation”. In: *arXiv preprint arXiv:1802.06955* (2018).

[20] Yao Qin et al. “Autofocus layer for semantic segmentation”. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part III 11*. Springer. 2018, pp. 603–611.

[21] Liang-Chieh Chen. “Rethinking atrous convolution for semantic image segmentation”. In: *arXiv preprint arXiv:1706.05587* (2017).

[22] Jiang-Jiang Liu et al. “A simple pooling-based design for real-time salient object detection”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 3917–3926.

[23] Hengshuang Zhao et al. “Pyramid scene parsing network”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2881–2890.

[24] Zaiwang Gu et al. “Ce-net: Context encoder network for 2d medical image segmentation”. In: *IEEE transactions on medical imaging* 38.10 (2019), pp. 2281–2292.

[25] Jo Schlemper et al. “Attention gated networks: Learning to leverage salient regions in medical images”. In: *Medical image analysis* 53 (2019), pp. 197–207.

[26] Xiaolong Wang et al. “Non-local neural networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 7794–7803.

[27] Wenyu Xing et al. “CM-SegNet: A deep learning-based automatic segmentation approach for medical images by combining convolution and multilayer perceptron”. In: *Computers in Biology and Medicine* 147 (2022), p. 105797.

[28] Alexey Dosovitskiy. “An image is worth 16x16 words: Transformers for image recognition at scale”. In: *arXiv preprint arXiv:2010.11929* (2020).

[29] Sixiao Zheng et al. “Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 6881–6890.

[30] Jieneng Chen et al. “Transunet: Transformers make strong encoders for medical image segmentation”. In: *arXiv preprint arXiv:2102.04306* (2021).

[31] Yundong Zhang, Huiye Liu, and Qiang Hu. “Transfuse: Fusing transformers and cnns 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. Springer. 2021, pp. 14–24.

[32] Shaohua Li et al. "Medical image segmentation using squeeze-and-expansion transformers". In: *arXiv preprint arXiv:2105.09511* (2021).

[33] Bingzhi Chen et al. "Transattunet: Multi-level attention-guided u-net with transformer for medical image segmentation". In: *IEEE Transactions on Emerging Topics in Computational Intelligence* (2023).

[34] Hong-Yu Zhou et al. "nnformer: Interleaved transformer for volumetric segmentation". In: *arXiv preprint arXiv:2109.03201* (2021).

[35] Andreas Ess et al. "Segmentation-Based Urban Traffic Scene Understanding." In: *BMVC*. Vol. 1. Citeseer. 2009, p. 2.

[36] Andreas Geiger, Philip Lenz, and Raquel Urtasun. "Are we ready for autonomous driving? the kitti vision benchmark suite". In: *2012 IEEE conference on computer vision and pattern recognition*. IEEE. 2012, pp. 3354–3361.

[37] Anping Xu et al. "Threshold-based level set method of image segmentation". In: *Intelligent Networks and Intelligent Systems, International Workshop on*. IEEE Computer Society. 2010, pp. 703–706.

[38] Cevahir Cigla and A Aydin Alatan. "Region-based image segmentation via graph cuts". In: *2008 15th IEEE International Conference on Image Processing*. IEEE. 2008, pp. 2272–2275.

[39] Zhao Yu-Qian et al. "Medical images edge detection based on mathematical morphology". In: *2005 IEEE engineering in medicine and biology 27th annual conference*. IEEE. 2006, pp. 6492–6495.

[40] Zhen Ma, João Manuel RS Tavares, and RM Natal Jorge. "A review on the current segmentation algorithms for medical images". In: *International conference on imaging theory and applications*. Vol. 1. SciTePress. 2009, pp. 135–140.

[41] Ana Ferreira, Fernanda Gentil, and João Manuel RS Tavares. "Segmentation algorithms for ear image data towards biomechanical studies". In: *Computer methods in biomechanics and biomedical engineering* 17.8 (2014), pp. 888–904.

[42] Zhen Ma et al. “A review of algorithms for medical image segmentation and their applications to the female pelvic cavity”. In: *Computer Methods in Biomechanics and Biomedical Engineering* 13.2 (2010), pp. 235–246.

[43] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[44] Christian Szegedy et al. “Going deeper with convolutions”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.

[45] Leonardo Rundo et al. “USE-Net: Incorporating Squeeze-and-Excitation blocks into U-Net for prostate zonal segmentation of multi-institutional MRI datasets”. In: *Neurocomputing* 365 (2019), pp. 31–43.

[46] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. “Segnet: A deep convolutional encoder-decoder architecture for image segmentation”. In: *IEEE transactions on pattern analysis and machine intelligence* 39.12 (2017), pp. 2481–2495.

[47] Jonathan Long, Evan Shelhamer, and Trevor Darrell. “Fully convolutional networks for semantic segmentation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3431–3440.

[48] Liang-Chieh Chen. “Semantic image segmentation with deep convolutional nets and fully connected CRFs”. In: *arXiv preprint arXiv:1412.7062* (2014).

[49] Liang-Chieh Chen et al. “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs”. In: *IEEE transactions on pattern analysis and machine intelligence* 40.4 (2017), pp. 834–848.

[50] Liang-Chieh Chen. “Rethinking atrous convolution for semantic image segmentation”. In: *arXiv preprint arXiv:1706.05587* (2017).

[51] Liang-Chieh Chen et al. “Encoder-decoder with atrous separable convolution for semantic image segmentation”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 801–818.

[52] Olaf Ronneberger, Philipp Fischer, and Thomas 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. Springer. 2015, pp. 234–241.

[53] Zongwei Zhou et al. “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*. Springer. 2018, pp. 3–11.

[54] Md Zahangir Alom et al. “Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation”. In: *arXiv preprint arXiv:1802.06955* (2018).

[55] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[56] Ozan Oktay et al. “Attention u-net: Learning where to look for the pancreas”. In: *arXiv preprint arXiv:1804.03999* (2018).

[57] Reza Azad et al. “Bi-directional ConvLSTM U-Net with densley connected convolutions”. In: *Proceedings of the IEEE/CVF international conference on computer vision workshops*. 2019, pp. 0–0.

[58] Shuanglang Feng et al. “CPFNet: Context pyramid fusion network for medical image segmentation”. In: *IEEE transactions on medical imaging* 39.10 (2020), pp. 3008–3018.

[59] Jieneng Chen et al. “Transunet: Transformers make strong encoders for medical image segmentation”. In: *arXiv preprint arXiv:2102.04306* (2021).

[60] Hugo Touvron, Matthieu Cord, and Hervé Jégou. “Deit iii: Revenge of the vit”. In: *European conference on computer vision*. Springer. 2022, pp. 516–533.

[61] Bo Dong et al. “Polyp-pvt: Polyp segmentation with pyramid vision transformers”. In: *arXiv preprint arXiv:2108.06932* (2021).

[62] Ze Liu et al. “Swin transformer: Hierarchical vision transformer using shifted windows”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 10012–10022.

[63] Enze Xie et al. “SegFormer: Simple and efficient design for semantic segmentation with transformers”. In: *Advances in neural information processing systems* 34 (2021), pp. 12077–12090.

[64] Zhendong Wang et al. “Uformer: A general u-shaped transformer for image restoration”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 17683–17693.

[65] Jieneng Chen et al. “Transunet: Transformers make strong encoders for medical image segmentation”. In: *arXiv preprint arXiv:2102.04306* (2021).

[66] Yundong Zhang, Huiye Liu, and Qiang Hu. “Transfuse: Fusing transformers and cnns 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*. Springer. 2021, pp. 14–24.

[67] Bingzhi Chen et al. “Transattunet: Multi-level attention-guided u-net with transformer for medical image segmentation”. In: *IEEE Transactions on Emerging Topics in Computational Intelligence* (2023).

[68] Junying Chen, Haijun You, and Kai Li. “A review of thyroid gland segmentation and thyroid nodule segmentation methods for medical ultrasound images”. In: *Computer methods and programs in biomedicine* 185 (2020), p. 105329.

[69] Walid Al-Dhabyani et al. “Dataset of breast ultrasound images”. In: *Data in brief* 28 (2020), p. 104863.

[70] Epimack Michael et al. “Breast cancer segmentation methods: current status and future potentials”. In: *BioMed research international* 2021.1 (2021), p. 9962109.

[71] Haifan Gong et al. “Thyroid region prior guided attention for ultrasound segmentation of thyroid nodules”. In: *Computers in biology and medicine* 155 (2023), p. 106389.

[72] David Gutman et al. “Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC)”. In: *arXiv preprint arXiv:1605.01397* (2016).

[73] Noel CF Codella 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)*. IEEE. 2018, pp. 168–172.

[74] Noel Codella et al. “Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic)”. In: *arXiv preprint arXiv:1902.03368* (2019).

[75] Teresa Mendonça et al. “PH 2-A dermoscopic image database for research and benchmarking”. In: *2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE. 2013, pp. 5437–5440.

[76] Marc Combalia et al. “Bcn20000: Dermoscopic lesions in the wild”. In: *arXiv preprint arXiv:1908.02288* (2019).

[77] Veronica Rotemberg et al. “A patient-centric dataset of images and metadata for identifying melanomas using clinical context”. In: *Scientific data* 8.1 (2021), p. 34.

[78] Bill Cassidy et al. “Analysis of the ISIC image datasets: Usage, benchmarks and recommendations”. In: *Medical image analysis* 75 (2022), p. 102305.

[79] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. “The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions”. In: *Scientific data* 5.1 (2018), pp. 1–9.

[80] Duwei Dai et al. “Ms RED: A novel multi-scale residual encoding and decoding network for skin lesion segmentation”. In: *Medical image analysis* 75 (2022), p. 102293.

[81] Xinzi He et al. “Fully transformer network for skin lesion analysis”. In: *Medical Image Analysis* 77 (2022), p. 102357.

[82] Ziyun Yang and Sina Farsiu. “Directional connectivity-based segmentation of medical images”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, pp. 11525–11535.

[83] Dhiraj Maji, Prarthana Sigedar, and Munendra Singh. “Attention Res-UNet with Guided Decoder for semantic segmentation of brain tumors”. In: *Biomedical Signal Processing and Control* 71 (2022), p. 103077.

[84] Hu Cao et al. “Swin-unet: Unet-like pure transformer for medical image segmentation”. In: *European conference on computer vision*. Springer. 2022, pp. 205–218.

[85] Ozan Oktay et al. “Attention u-net: Learning where to look for the pancreas”. In: *arXiv preprint arXiv:1804.03999* (2018).

[86] Razvan-Gabriel Dumitru, Darius Peteleaza, and Catalin Craciun. “Using DUCK-Net for polyp image segmentation”. In: *Scientific reports* 13.1 (2023), p. 9803.

[87] Quoc-Huy Trinh. “Meta-Polyp: a baseline for efficient Polyp segmentation”. In: *2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE. 2023, pp. 742–747.

[88] Huisi Wu et al. “FAT-Net: Feature adaptive transformers for automated skin lesion segmentation”. In: *Medical image analysis* 76 (2022), p. 102327.

[89] Kai Hu et al. “AS-Net: Attention Synergy Network for skin lesion segmentation”. In: *Expert Systems with Applications* 201 (2022), p. 117112.

[90] Tsung-Yi Lin et al. “Feature pyramid networks for object detection”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2117–2125.

[91] Baiying Lei et al. “Segmentation of breast anatomy for automated whole breast ultrasound images with boundary regularized convolutional encoder–decoder network”. In: *Neurocomputing* 321 (2018), pp. 178–186.

[92] Liang-Chieh Chen et al. “Encoder-decoder with atrous separable convolution for semantic image segmentation”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 801–818.

[93] Yunzhu Wu et al. “BGM-Net: boundary-guided multiscale network for breast lesion segmentation in ultrasound”. In: *Frontiers in Molecular Biosciences* 8 (2021), p. 698334.

[94] Raghav Mehta and Jayanthi Sivaswamy. “M-net: A convolutional neural network for deep brain structure segmentation”. In: *2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017)*. Ieee. 2017, pp. 437–440.

[95] Baiying Lei et al. “Skin lesion segmentation via generative adversarial networks with dual discriminators”. In: *Medical Image Analysis* 64 (2020), p. 101716.

[96] Lei Bi, Michael Fulham, and Jinman Kim. “Hyper-fusion network for semi-automatic segmentation of skin lesions”. In: *Medical image analysis* 76 (2022), p. 102334.

[97] Xingqing Nie et al. “N-Net: a novel dense fully convolutional neural network for thyroid nodule segmentation”. In: *Frontiers in Neuroscience* 16 (2022), p. 872601.

[98] Xiaoliang Jiang et al. “SEACU-Net: Attentive ConvLSTM U-Net with squeeze-and-excitation layer for skin lesion segmentation”. In: *Computer methods and programs in biomedicine* 225 (2022), p. 107076.

[99] Weiwei Cao et al. “ICL-Net: Global and local inter-pixel correlations learning network for skin lesion segmentation”. In: *IEEE Journal of Biomedical and Health Informatics* 27.1 (2022), pp. 145–156.

[100] Qing Xu et al. “DCSAU-Net: A deeper and more compact split-attention U-Net for medical image segmentation”. In: *Computers in Biology and Medicine* 154 (2023), p. 106626.

[101] Yanjun Peng, Dian Yu, and Yanfei Guo. “MShNet: Multi-scale feature combined with h-network for medical image segmentation”. In: *Biomedical Signal Processing and Control* 79 (2023), p. 104167.