Development of Attention Guided U-Net for Medical Image Segmentation

Subrato Bharati

A Thesis

in

The Department

of

Electrical and Computer Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Master of Applied Science (Electrical and Computer Engineering) at

Concordia University

Montréal, Québec, Canada

August 2025

CONCORDIA UNIVERSITY
SCHOOL OF GRADUATE STUDIES

This is to certify that the thesis prepared

By: **Subrato Bharati**

Entitled: **Development of Attention Guided U-Net for Medical Image Segmentation**
and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science (Electrical and Computer Engineering)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

_____ Chair
Dr. Chunyan Wang

_____ External Examiner
Dr. Chun-Yi Su (MIAE)

_____ Internal Examiner
Dr. Chunyan Wang

_____ Supervisor
Dr. M. Omair Ahmad

_____ Supervisor
Dr. M.N.S. Swamy

Approved by _____
Dr. Abdelwahab Hamou-Lhadj, Chair
Department of Electrical and Computer Engineering

_____ 2025 _____
Dr. Mourad Debbabi, Dean
Gina Cody School of Engineering and Computer Science

Abstract

Development of Attention Guided U-Net for Medical Image Segmentation

Subrato Bharati

Medical image segmentation is a process of isolating or identifying an object of interest in medical images. It plays a pivotal role in clinical diagnostics, monitoring and treating diseases. The convolutional neural network, U-Net, was specifically developed for segmentation of medical images in view of its ability to accurately segment with limited training data. Existing U-Net based segmentation networks suffer from high computational complexity in order to provide a reasonable performance. This thesis presents five U-Net based schemes that significantly reduce the computational complexity without compromising the performance. In the first part, we develop a number of segmentation schemes referred to as MAGNet, MedSegNet, SSNet, and FFNet that utilize attention mechanism-enhanced multiscale feature fusion. The first two networks are developed for segmenting CT, colonoscopy, and non-mydriatic 3CCD images. SSNet is a semi-supervised network that effectively makes use of both labeled and unlabeled data for segmenting brain anatomical structures of tissues in MR images. FFNet is proposed for segmenting benign and malignant tumors from ultrasound images. In the second part, we present a lightweight attention-guided network with feature recalibration, referred to as LASegNet. The main idea used in designing LASegNet is on one hand to reduce the number of parameters by cutting on the number of filters used and on the other hand, restore the performance by combining the features of the encoder and decoder units through a judicious use of attention guided module. Extensive

experiments are performed to demonstrate the effectiveness of each of the schemes proposed. Specifically, it is shown that LASegNet is robust across images from different modalities.

Acknowledgements

I wholeheartedly extend my gratitude to my thesis supervisors, Professor Dr. M. Omair Ahmad and Professor Dr. M.N.S. Swamy, who have been exceptional supervisors throughout my journey. Their constant motivation, proper guidance, and invaluable insights have been instrumental in shaping this thesis. Their unwavering support and creative ideas have consistently pushed my abilities to new heights and inspired me to aim for excellence. I am thankful to my supervisors, whose discussions and suggestions have played a pivotal role in the progression of this work. Their constructive criticisms have helped enhance the quality and rigor of this research. I also thank my university and the academic community for their support in completing this thesis. I would also like to express my sincere gratitude to Professor Dr. M. Rubaiyat Hossain Mondal for his continuous motivation and support throughout the journey.

To my dear wife, Mithila Das, your love, encouragement, and sacrifices have been the backbone of this journey. Your unwavering belief in me has been a source of strength during challenging times, and I cannot thank you enough for standing by my side.

Finally, I extend my heartfelt gratitude to my family and friends for their endless support, inspiration, and understanding. Your encouragement has been the foundation of my accomplishments, and I am forever grateful for the role you have played in this milestone.

Contents

List of Figures

List of Tables

List of Abbreviations

Adam	Adaptive Moment Estimation
AdamW	Adam with Decoupled Weight Decay
AG	Attention Gate
ASPP	Atrous Spatial Pyramid Pooling
BCE	Binary Cross-Entropy
BN	Batch Normalization
BUS	Breast Ultra Sound
BUSI	Breast UltraSound Images
CDL	Contextual Differential Loss
CNNs	Convolutional Neural Networks
CSA	Channel Self-Attention
CSGM	Contour-aware Super-pixel Grid Mixing
CT	Computed Tomography
DDTI	Digital Database for Thyroid Imaging

DL	Deep Learning
DSA	Dual Self-Attention
DSC	Dice Similarity Coefficient
EM	Expectation-Maximization
fAdamR	Fusion of Adam and RMSProp
FC	Fully Connected
FCM	Fuzzy C-Means
FFNet	Feature Fusion-based Network
GAN	Global Attention Network
HFIM	High-level Feature Improvement Module
IoU	Intersection over Union
LASegNet	Lightweight Attention-guided Segmentation Network
MAGNet	Multi-scale and Global Attention Network
MRI	Magnetic Resonance Imaging
MSA	Multi-Scale Attention
MSE	Mean Squared Error
PSA	Position Self-Attention
ReLU	Rectified Linear Unit

RMSProp	Root Mean Square Propagation
SE	Squeeze-and-Excitation
SGD	Stochastic Gradient Descent
SGL	Self-Guided Learning
SSNet	Semi-Supervised Network
Std	Standard Deviation

List of Symbols

β_1	First order moment decay
β_2	Second order moment decay
C	Number of channels
$Conv$	Convolution operation
D_{CM}	The contextual mapping of the images
ε	Smoothing term
F	Feature Map
H	Height
K	Key
$\mathcal{L}_{dice}$	Dice loss
$\mathcal{L}_{MSE}$	MSE loss between the predicted displacement field and the ground truth displacement field
η	Learning rate
Q	Query

R_{pred}	Predicted displacement field
R_{true}	Ground truth displacement field
S_{Mix}	Superpixel map
S_{pred}	Predicted segmentation
S_{true}	Ground truth segmentation
V	Value
W	Width
w_i	The weight assigned to the *i*-th pixel
X_{Mix}	Augmented image
Y_{Mix},	Augmented corresponding mask
y_i	Ground truth label
$\widehat{y_i}$	Predicted probability
λ	The weight decay coefficient
$\odot$	Element wise multiplication

Chapter 1

Introduction

1.1 General

The primary goal of medical image segmentation is to identify and isolate those areas of a medical image that are affected by a disease [1]. It plays a crucial role in a variety of clinical applications including diagnosis, treatment planning, surgical navigation, and disease monitoring by enabling precise delineation and quantification of anatomical structures from different imaging modalities such as Magnetic Resonance Imaging (MRI), Computed Tomography (CT), Ultrasound, and endoscopy images. One of the significant challenges in automatic segmentation of a medical image as opposed to that of a natural image is the inherent differences between medical and natural images, and medical images vary in terms of textures, shapes [2].

Traditional segmentation schemes typically depend on handcrafted features, mathematical models, and predefined rules to identify regions of interest or objects of interest based on intensity, texture, or shape [3]. These traditional schemes struggle with the different modalities of medical images due to their reliance on handcrafted features, which makes them less adaptable to the diverse variations in shape, texture, and noise commonly found in real-world medical images. One of the simplest schemes for medical image segmentation involves intensity thresholding, where each

pixel is classified based on a predefined intensity threshold obtained from the histogram of the image. This approach compares the intensity of each pixel to the selected threshold value to assign it to a corresponding class. On the other hand, while the traditional method is computationally efficient and straightforward, it lacks the ability to incorporate spatial context, which makes it highly sensitive to noise, intensity variations, and the selection of the threshold value itself [3]. To address these limitations, more advanced techniques, such as clustering and region growth, have been introduced, which incorporate neighborhood information and indicate robustness. The region growing method starts with a set of user-defined seed points and incrementally aggregates neighboring pixels that satisfy a similarity criterion, such as minimal intensity difference with respect to the pixels of seed. Despite its effectiveness in capturing homogeneous regions, this traditional method remains sensitive to the choice of seed points and intensity (non-uniformity) and often requires manual initialization [4]. Further, clustering-based methods provide an unsupervised alternative by grouping pixels with similar properties without requiring explicit seed point selection. These algorithms typically begin with randomly initialized cluster centers and iteratively refine pixel groupings based on statistical similarity. Prominent clustering techniques applied in medical image segmentation include Fuzzy C-Means (FCM), k-Means, and Expectation-Maximization (EM), which are capable of handling intensity inhomogeneity through iterative refinement and correction strategies [5].

Deep learning (DL)-based technique has recently emerged as a dominant approach in biomedical image segmentation [6]. DL-based techniques enable automated segmentation across various types of biomedical images, which leverage large-scale data to automatically learn hierarchical features, enabling more robust and accurate segmentation without the need for extensive manual feature engineering [1]. Over the years, numerous DL-based segmentation models have been developed to

enhance accuracy and efficiency in medical image segmentation. Medical image segmentation has seen rapid progress with the advent of deep convolutional neural networks (CNNs), particularly encoder-decoder architectures. Over the past decade, many researchers have developed specialized models for different medical imaging modalities to address challenges of low contrast, complex textures, variable object size, and class imbalance. We now review DL-based prominent networks categorized by imaging modality and analyze their strengths and limitations, ultimately motivating the adoption of U-Net-based architectures and their enhancements.

Computed tomography (CT) images are widely used in segmenting organs and pathological regions due to their high-resolution cross-sectional views. Several architectures have been proposed for the segmentation of an entire organ or a certain region of abnormalities from CT scan images. The authors of [7] proposed Inf-Net, which was developed for COVID-19 lung infections segmentation and introduced an implicit reverse attention mechanism to capture local features, but it suffers from instability in low-contrast boundaries. The work of [8] introduced Gated UNet, which integrated spatial and channel attention gates into the U-Net architecture, enhancing focus on salient features, though it increases model complexity and inference time. Another work, DeepLab3+ [9], with its atrous spatial pyramid pooling (ASPP), captured multi-scale context effectively but might struggle with accurate edge localization in CT images. The authors of [10] proposed DenseUNet, which combined dense connectivity with the structure of U-Net, improved gradient flow, and feature reuse but introduced the computational cost and required an amount of memory for large volumes of CT scans.

Polyp segmentation in colonoscopy or endoscopy images is a challenging task due to their shapes, textures, and illumination. The authors of [11] developed ColonSegNet, which offered a lightweight architecture tailored for real-time segmentation; however, ColonSegNet was less

effective on challenging polyps with indistinct boundaries. Later, authors of [12] proposed PraNet, which introduced reverse attention modules and boundary-aware supervision, significantly improving delineation of polyp edges. However, its reliance on a hand-crafted attention mechanism can limit the generalizability of PraNet. The authors of [13] applied DeepLab3+ for polyp segmentation, though its performance was often surpassed in different modalities of medical images.

Retinal blood vessel segmentation is vital for diagnosing diseases such as diabetic retinopathy and hypertension. The authors of [14] employed CS-Net, which combined contextual and spatial features, enabling better performance on small vascular structures, but it can be sensitive to noise in low-resolution images. Next, the authors of [15] introduced a network called Self-Guided Learning (SGL), which presented a curriculum learning strategy for vessel segmentation, improving convergence but requiring schedule tuning. The authors of [16] proposed a GAN-based approach, RV-GAN, which enhanced realism in the prediction of segmentation but may introduce artifacts and instability during training. Another network called Attention UNet [17] and FANet [19] both utilized attention mechanism and FR-UNet [18] utilized residual mechanism to emphasize vascular regions, yet their performance was heavily influenced by the quality of retinal fundus image preprocessing.

Ultrasound image segmentation is a complicated task due to its low contrast, speckle noise, and poor boundary definition. The authors of [9] developed a variant of Modified DeepLab3+ with attention or edge modules that was used to tackle these issues, but performance was limited while dealing with complex anatomical structures. Later, the authors of [90] proposed TransUNet, which integrated a transformer-based encoder into U-Net, captures long-range dependencies effectively but requires large datasets and high computation. An another network, Attention UNet [20], has

shown promise in handling fuzzy boundaries through spatial attention, though the design can be sensitive to initialization.

Brain MR image segmentation is essential for analyzing tumors, lesions, and brain structures. A classic encoder-decoder model, SegNet [59], was applied for brain tissue segmentation, but it lacks skip connections, which limits its ability to recover fine-grained spatial details. DeepLab-based architecture, DeepLabv3+ [58], helped to incorporate multi-scale information but often underperformed near region boundaries in complex brain MR scans.

Based on their architecture, segmentation networks can be broadly grouped into three categories. The first category includes well-known encoder-decoder architectures, U-Net [21], Attention U-Net [20], Gated U-Net [8], and FRUNet [18], which use hierarchical upsampling layers and skip connections to fuse semantic and spatial information. These networks have demonstrated strong performance across various medical imaging modalities, including CT, retinal fundus images, and ultrasound. However, the introduction of attention modules or deep skip connections often leads to an increase in parameter count and memory consumption in these networks. Additionally, while skip connections help preserve spatial details, these networks may also pass irrelevant low-level features, leading to noisy outputs in complex or low-contrast images. The second category focuses on multi-scale and context-aware architectures, DeepLabv3+ [9], DenseUNet [10], ColonSegNet [11], and PraNet [12]. These multi-scale and context-aware networks use some techniques like atrous spatial pyramid pooling (ASPP), dense connectivity, or boundary refinement to aggregate information at multiple scales. These networks have shown strong performance in segmenting challenging structures of polyps in endoscopic images or lung infections in CT scans. Despite their robustness in handling size and shape variability, these networks often rely on heavy backbone networks or large receptive fields, which increase computational cost and reduce spatial accuracy,

especially in edge-localized tasks. The third category includes transformer-based networks or hybrid networks, TransUNet [90], SGL [15], and RV-GAN [16], which integrated non-local attention or transformer encoders to capture global features. These networks are particularly useful in complex views of retinal vessel segmentation or segmenting abnormalities from ultrasound images, where local pixel-based features alone are insufficient. While effective in global context modeling, typically, these architectures are computationally expensive and sensitive to hyperparameter tuning, which can impact the deployment in real-time clinical applications. Regardless of the diversity of these architectures, some common limitations persist across modalities and tasks. In CT and MRI images, Inf-Net [7] and DeepLabv3+ [9] may struggle with low-contrast boundaries. In endoscopic images, although PraNet [12] and ColonSegNet [11] achieve accurate segmentation of polyps, these networks may underperform in the presence of flat or small lesions. For retinal images, attention-based networks, FRUNet [18] and CS-Net [14], often face challenges in capturing the finest vascular structures under illumination variations. In ultrasound imaging, TransUNet [90] and Attention UNet [20] provide notable improvements but require large amounts of data.

Most of the above architectures have been developed for specific imaging modalities; most still struggle with modality-specific challenges such as boundary ambiguity, noise, and class imbalance. The U-Net architecture has become a foundational model due to its simplicity and effectiveness, and it has been employed for the segmentation of the images of different modalities. The symmetric encoder-decoder structure with skip connections helps preserve high-resolution spatial features, which is crucial for isolating small or complex anatomical structures. Despite its advantages, U-Net suffers from limitations such as fixed receptive fields, limited global context modeling, and performance degradation in highly imbalanced datasets. To address these limitations, several

enhancements have been proposed, including attention mechanisms, such as Attention UNet [20] and Gated U-Net [8], multi-scale modules, such as DeepLabv3+ [9], and hybrid networks that include TransUNet [90], MISSFormer [91]. These networks improve feature discrimination, contextual awareness, and structural consistency of the predicted segmentation masks. In summary, U-Net and its variants offer a strong baseline, and further improvements are possible through the integration of attention modules and hybrid encoder designs. The design and analysis of such improvements constitute the core motivation of this thesis.

1.2 Motivation

These above observations highlight the need for a segmentation architecture that combines the efficiency and performance of U-Net, the contextual awareness of multi-scale networks, and the discriminative power of attention mechanisms, while maintaining a lightweight structure suitable for clinical deployment. By integrating adaptive attention modules, improved feature fusion strategies, and carefully designed loss functions, it is possible to overcome the limitations of existing networks, enhancing boundary precision, reducing computational cost, and increasing robustness to modality-specific challenges. The motivation of this thesis is thus rooted in addressing these gaps by proposing a U-Net-based architecture that is both computationally efficient and capable of delivering high segmentation accuracy across multiple imaging modalities.

1.3 Objective

Motivated by the limitations identified in current state-of-the-art segmentation networks with respect to computational complexity, poor boundary localization, isolation of an entire organ or a

certain region of abnormalities within the organ, and poor performance across heterogeneous datasets, this thesis aims to design a U-Net-based architecture enhanced with attention mechanisms and multi-scale feature for medical image segmentation of various modalities medical images, including CT scan, endoscopy, ultrasound, and retinal fundus images. To address modality-specific challenges, the proposed networks incorporate several key novelties: (1) a lightweight encoder-decoder structure with integrated spatial and channel-wise attention to capture both local and global features; (2) customized data augmentation strategies to reduce the overfitting of the network; (3) a hybrid loss function to balance region overlap and boundary accuracy, particularly in class-imbalanced settings; and (4) optimized training protocols using adaptive optimizers to achieve a high performance for different datasets. The performance of the proposed network is extensively evaluated with that of the state-of-the-art networks in terms of standard metrics.

1.4 Organization of the Thesis

This thesis is organized as follows. Chapter 2 presents the essential background material relevant to this work, including an overview of the U-Net architecture, attention mechanisms, loss functions, optimizers, and evaluation metrics commonly used in medical image segmentation. Chapter 3 develops two proposed architectures, MAGNet and MedSegNet, for segmenting lung infection, polyp, and retinal blood vessels, respectively, from CT, colonoscopy, and non-mydriatic 3CCD images. Chapter 3 also develops a semi supervised network for the segmentation brain tissue to diagnose abnormalities in the brain from brain MRI, simultaneously addressing spatial alignment and tracking disease progression. Finally, this chapter also presents a novel framework for breast cancer segmentation in ultrasound images, utilizing super-pixel grid mixing-based augmentation, a contextual differential loss function, and a feature fusion network. Chapter 4 provides a lightweight

attention-guided segmentation network that incorporates feature recalibration for segmentation of different modality medical images. Chapter 5 concludes the thesis by summarizing the key contributions, and providing directions for future research.

Chapter 2

Background Material

2.1 Introduction

The goal of this chapter is to provide an overview of the foundational concepts and methodologies that support the research presented in this thesis. As deep learning continues to play a transformative role in medical image analysis, understanding its core components is essential for both the development and evaluation of robust segmentation models. This chapter outlines the theoretical and practical background required to contextualize the proposed methods and to interpret their performance in subsequent chapters. Specifically, the chapter begins with an introduction to U-Net, which forms the backbone of most state-of-the-art architectures used in medical image segmentation. Their structural design, feature extraction capabilities, and hierarchical learning mechanisms are discussed in detail to establish a solid understanding of their functionality. Furthermore, the chapter delves into a thorough examination of commonly utilized loss functions, emphasizing their role in guiding model optimization in segmentation tasks. The chapter illustrates loss functions like cross-entropy and dice loss, highlighting their theoretical formulation and practical implications. Finally, the chapter introduces key evaluation metrics that are used to quantitatively assess segmentation performance. Metrics such as Dice Similarity

Coefficient (DSC), Intersection over Union (IoU), sensitivity, and specificity are described, highlighting their relevance in validating clinical applicability and benchmarking against existing approaches. Together, these foundational topics establish the necessary background for understanding the design, training, and evaluation of deep learning-based medical image segmentation systems, and they form the basis for the novel contributions presented in later chapters.

2.2 U-Net Architecture

A convolutional neural network, U-Net, was originally proposed by Ronneberger et al. [21] to address semantic segmentation tasks. Characterized by its symmetric U-shaped design, referred to in Figure 2.1, the U-Net model comprises two primary components, a feature extracting path on the left and a feature reconstructing path on the right. The feature extracting path, also called the encoder, performs successive convolutional operations and downsampling operations to extract hierarchical features from the input image. This stage is responsible for compressing high-dimensional input data into a compact latent representation that captures features of the image. Conversely, the feature reconstructing, or decoder, progressively upsamples and refines these encoded features to restore the spatial resolution necessary for pixel-wise classification. To enhance reconstruction in the decoding path, skip connections are introduced between corresponding layers of the encoder and decoder, enabling the network to retain high-resolution contextual information. The last layer of the U-Net employs a 1×1 convolution to provide the output feature maps with the desired number of segmentation classes. The effectiveness of the encoder in capturing discriminative features has a direct impact on the overall performance of the network [22]. Accordingly, the subsequent sections of this thesis provide an in-depth overview of

advanced encoder architectures integrated within various segmentation networks explored in this work.

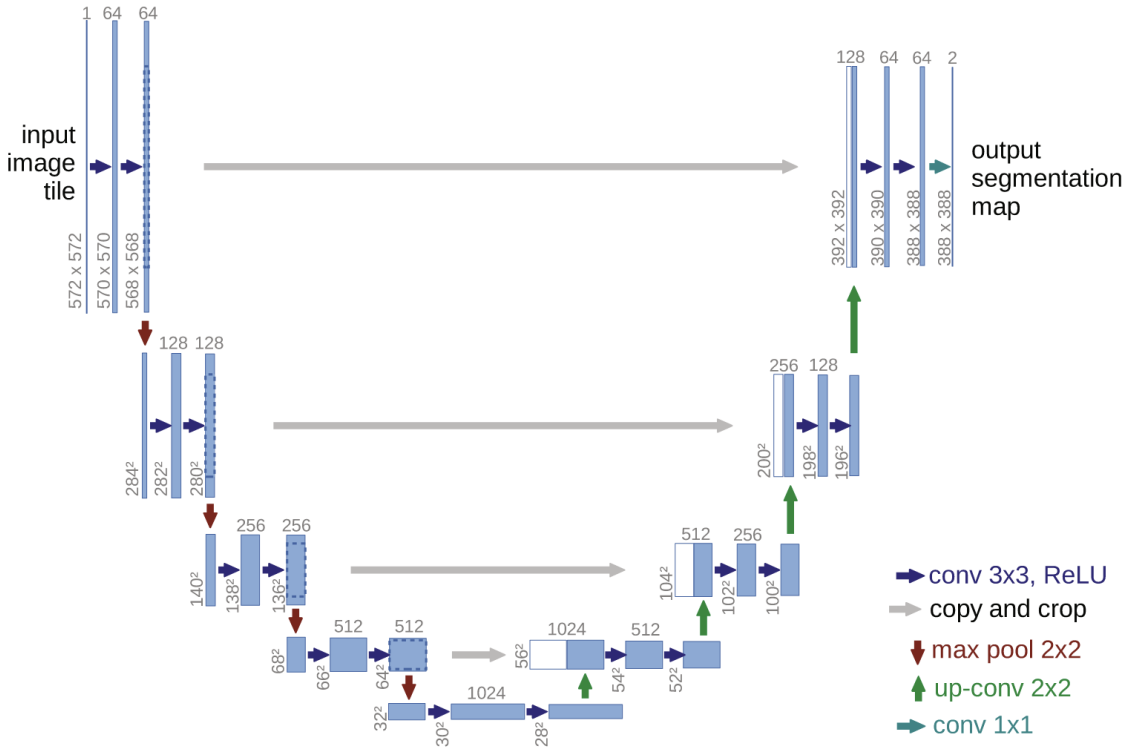


Figure 2.1: U-Net architecture [21]

2.3 Attention Mechanism

The attention mechanism is a computational strategy originally developed in the context of natural language processing to allow models to selectively focus on relevant parts of the input sequence when generating predictions [46]. In recent years, attention mechanisms have been successfully adapted to computer vision and, more specifically, to medical image segmentation [23]. By enabling networks to dynamically highlight informative features while suppressing irrelevant ones, attention modules improve the capacity of the model to handle complex anatomical variations, blurred boundaries, and subtle pathological patterns commonly found in medical images.

Convolutional Neural Networks (CNNs), including U-Net and its variants, form the backbone of many state-of-the-art segmentation networks. These architectures rely on stacked convolutional layers to extract features at multiple scales. However, conventional convolutions are inherently local operators, limited by their receptive fields, and lack the ability to capture long-range dependencies or global contextual relationships that are critical for accurate segmentation. Medical images, CT, MRI, or ultrasound, often contain pathologies that may not be easily distinguished alone based on local pixel intensity. For example, lesions with irregular shapes, occlusions, or intensity inhomogeneity require global awareness for accurate delineation. Attention modules address this challenge by allowing the model to consider the importance of different spatial or channel-wise features, effectively learning where and what to focus on. In medical image segmentation, the attention modules are commonly inserted into the skip connections or bottleneck layers of U-Net or its variants. There are two primary types of attention mechanisms used, such as spatial attention, which determines where to focus on the image by generating attention maps across spatial dimensions, and channel attention, which determines what features are important for the task by recalibrating feature maps across the channel dimension. Note that the attention U-Net enhances the performance of the original U-Net by incorporating gating-based attention blocks to refine skip connection features before merging them with the decoder. As an example, the scaled dot-product attention module is shown in Figure 2.2.

Let the input feature map extracted by a CNN be denoted as $X \epsilon R^{H \times W \times C}$, where H and W are the spatial dimensions height and width. C is the number of channels which produce features. Query (Q), Key (K), and Value (V) are computed by applying learnable weight from this feature map X. The attention module first projects this input into three learned representations:

$$Q = XW_Q, K = XW_K, V = XW_V \tag{2.1}$$

wherein $W_Q, W_K, W_V \epsilon R^{C \times d}$ are the learnable projection matrices, and d is the attention embedding dimension. The attention scores are computed using scaled dot-product attention by

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{D}})V \quad (2.2)$$

wherein $\sqrt{D}$ is the scaling factor to prevent exploding gradients (stabilizes training) and QK^T is the dot product similarity between all queries and keys. In attention, Q, K, and V are three projections of the image features, where Query (Q) asks what context a pixel or patch needs, Key (K) provides reference features, and Value (V) carries the actual information. This mechanism allows each spatial location in the image to gather information from all other locations, guided by the similarity between query and key vectors. The output is a context-enhanced representation that captures both local and global features. Attention modules enhance the representational power of convolutional networks by introducing dynamic weighting mechanisms that prioritize important features based on their relevance to the segmentation task. Their integration into architectures like U-Net has demonstrated consistent improvements in both quantitative metrics and qualitative segmentation quality, making them efficient in the domain of medical image analysis.

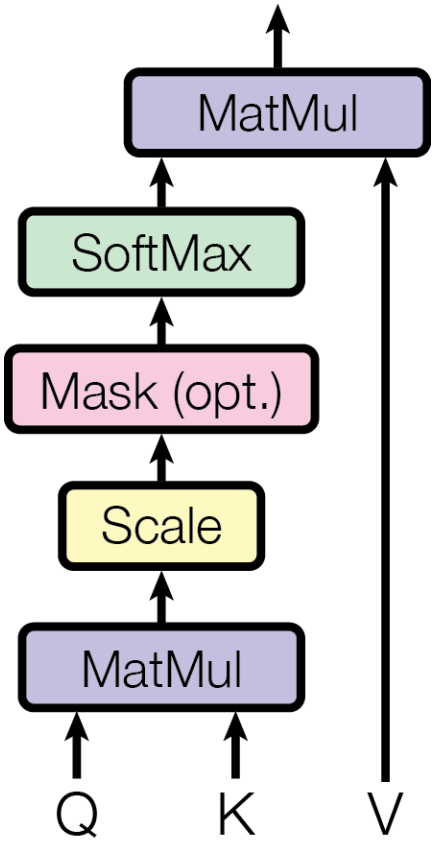


Figure 2.2: Scaled dot-product attention module [23]

2.4 Loss Functions

A wide range of loss functions have been proposed for medical image segmentation tasks; however, two of the most commonly employed are the Binary Cross-Entropy (BCE) loss [24] and the dice loss [25]. These loss functions play a critical role in guiding the network during training by calculating the difference between the predicted segmentation map and the ground truth annotation. Since a loss function quantifies how far the predicted segmentation is from the ground truth, guiding the network to update weights. Different loss functions are used in different network because medical images face challenges like class imbalance, small lesions, and blurry boundaries. Choosing the right loss ensures the network emphasizes clinically relevant structures. The binary

cross-entropy loss is derived from information theory and is widely used in binary classification problems. In the context of image segmentation, it treats each pixel as an independent binary classification task. For a predicted probability $\hat{y}_i \epsilon [0,1]$ and corresponding ground truth label $y_i \epsilon \{0, 1\}$, the BCE loss for a single pixel is given by

$$\mathcal{L}_{BCE} = -(y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)) \tag{2.3}$$

Combined over all pixels in an image, the BCE loss evaluates how well the network distinguishes between foreground and background regions. While effective, BCE can be sensitive to class imbalance, which is often present in medical segmentation tasks, where foreground structures like tumors or lesions occupy a small fraction of the image.

Dice loss is another frequently employed loss function. Based on the Dice Similarity Coefficient (DSC), it directly measures the overlap between the predicted segmentation and the ground truth. The dice loss is defined by

$$\mathcal{L}_{Dice} = 1 - \frac{2 \sum_i \hat{y}_i y_i + \varepsilon}{\sum_i \hat{y}_i + \sum_i y_i + \varepsilon} \tag{2.4}$$

where ε is a small constant added to prevent division by zero. Unlike BCE, dice loss is particularly effective in handling class imbalance, as it maximizes the intersection over the union of the predicted and actual ground truth regions.

2.5 Optimizers

Various optimization algorithms have been developed and adopted to train deep learning models for medical image segmentation. Among them, the most commonly used optimizers are Adam [26], AdamW [27], and RMSProp [28], [29], because of their robustness, efficiency, and ability to adaptively adjust learning rates during training. These optimizers play a crucial role in minimizing the loss by updating model parameters in the direction that reduces prediction errors. The Adaptive

Moment Estimation (Adam) optimizer is one of the most widely used methods in deep learning. It combines the advantages of two other extensions of Stochastic Gradient Descent (SGD) [30] and AdaGrad [31]. Adam computes adaptive learning rates for each parameter by maintaining exponentially decaying averages of past gradients (first moment) and squared gradients (second moment). The parameter update rule is given by

$$\theta_t = \theta_{t-1} - \eta . \frac{\widehat{m_t}}{\sqrt{\widehat{v_t}} - \varepsilon} \tag{2.5}$$

wherein η is the learning rate, and $\widehat{m_t}$ and $\sqrt{\widehat{v_t}}$ are the bias-corrected first and second moment estimates, respectively. ε is a small constant to avoid division by zero. Adam is well-suited for segmentation tasks due to its fast convergence and ability to handle sparse gradients and non-stationary objectives.

The AdamW optimizer [27] is a variation of Adam that decouples weight decay used for regularization from the gradient update. Unlike the Adam optimizer, where L2 regularization is intertwined with the moment estimation, AdamW applies weight decay in a more principled and independent way, as given by

$$\theta_t = \theta_{t-1} - \eta \left(\frac{\widehat{m_t}}{\sqrt{\widehat{v_t}} - \varepsilon} + \lambda \theta_{t-1} \right) \tag{2.6}$$

where λ is the weight decay coefficient. AdamW often results in better generalization, especially in models prone to overfitting, and is particularly beneficial in medical imaging where datasets are often limited.

The Root Mean Square Propagation (RMSProp) optimizer [28], [29] is another adaptive learning rate method that divides the gradient by a moving average of its recent magnitudes. It helps to stabilize and accelerate convergence by normalizing updates as follows.

$$\theta_t = \theta_{t-1} - \eta . \frac{g_t}{\sqrt{E[g^2]_t} - \varepsilon} \tag{2.7}$$

where $\sqrt{E[g^2]_t}$ is the exponentially weighted moving average of squared gradients. RMSProp is particularly effective in handling optimization problem and is known for its stability in training segmentation networks. The choice of optimizer depends on the specific architecture, characteristics of dataset, and computational resources. However, Adam and its variant AdamW are generally favored in medical image segmentation tasks for their adaptability and ease of hyperparameter tuning.

2.6 Performance Metrics

In this section, several performance metrics that are employed to evaluate the performance of the various segmentation networks introduced in this thesis are described below.

Intersection over Union (IoU)

IoU [32] measures the overlap between the predicted segmentation and the ground truth. It is defined as the ratio of the intersection area to the union area of the predicted and ground truth masks as follows.

$$IoU = \frac{TP}{FP+FN+TP} \tag{2.8}$$

Mean IoU also indicates the average IoU across all classes and provides a comprehensive measure of overlap. However, it can penalize small errors in medical images, where even slight inaccuracies may be critical.

Dice Coefficient (Dice Similarity Coefficient or DSC)

Dice coefficient [33] measures the similarity between the predicted segmentation and the ground truth. It is calculated by

$$DSC = \frac{2 \times TP}{FP + FN + 2 \times TP} \tag{2.9}$$

Dice coefficient is related to IoU but gives more weight to overlap and is often preferred in medical segmentation tasks due to its sensitivity to imbalances in class distributions. A high dice score indicates excellent segmentation performance. DSC is generally more sensitive to small object segmentation than IoU.

Precision

Precision [34] measures the proportion of true positive predictions relative to all positive predictions. Precision is calculated by

$$Precision = \frac{TP}{FP + TP} \tag{2.10}$$

Precision is particularly important when false positives are more critical than false negatives.

Recall or Sensitivity

Recall [34] measures the proportion of true positive predictions relative to all actual positives in the ground truth. Recall is calculated by

$$Recall = \frac{TP}{FN + TP} \tag{2.11}$$

High recall ensures the model captures as many true positives as possible, which is crucial in detecting anomalies or abnormalities.

Accuracy

Accuracy [35], [36] measures the proportion of correctly classified pixels (both true positives and true negatives) relative to the total number of pixels. Accuracy is calculated by

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \tag{2.12}$$

Accuracy [36] is a straightforward metric but can be misleading in medical segmentation due to class imbalance (e.g., background pixels significantly outnumber foreground pixels).

2.7 Summary

This chapter has presented the essential background concepts that are necessary to understand the research work presented in this thesis. The U-Net architecture has been introduced as a widely adopted encoder-decoder framework capable of learning both local and global image features, making it highly suitable for segmentation tasks. To further enhance spatial and contextual understanding, the concept of the attention module has been discussed, emphasizing its role in enabling the network to pay more attention to the important features. In addition, we have also presented some useful loss functions such as binary cross-entropy and dice loss, which are essential for guiding the model toward accurate segmentation. Popular optimization algorithms, including Adam, AdamW, and RMSProp, have also been presented, highlighting their significance in efficient and stable training of deep neural networks. Finally, standard performance metrics used for evaluating the segmentation performance have been described.

Chapter 3

Attention Mechanism Enhanced U-Net based Multiscale Feature Fusion Networks for Medical Image Segmentation

3.1 Introduction

In medical image segmentation tasks involving complex anatomical structures across different modality medical images, the ability of U-Net to generate rich multi-scale representations and network spatial dependencies is critical for achieving high accuracy. Since medical images often exhibit variability in shape, texture, and boundary definition, especially across modalities such as CT, colonoscopy, retinography, ultrasound, and brain MRI, a segmentation network with strong attention mechanisms and multi-scale feature fusion becomes essential for enhancing robustness and anatomical consistency of a network.

In this chapter, we propose and evaluate four different U-Net based architectures that employ different attention strategies, feature fusion methods, and training enhancements to address the

diverse challenges of medical image segmentation. Specifically, in Section 3.2, we propose a deep convolutional network with multi-scale and global attention modules, MAGNet [37], that enhances feature encoding across three distinct imaging modalities. In Section 3.3, we propose a network, MedSegNet [38], that integrates residual connections, dual self-attention, and multi-scale attention to achieve computational efficiency over that of MAGNet without compromising the segmentation performance. In section 3.4, we propose a multi-scale attention-enhanced semi-supervised network, SSNet, for segmentation of brain MRI images [39]. In Section 3.5, we propose a convolutional neural network with a multipath encoder, FFNet, that utilizes a data augmentation strategy called super-pixel grid mixing and a specialized contextual differential loss for breast cancer segmentation of ultrasound images [40]. Experiments are carried out for each of the proposed networks to evaluate its performance and the results compared with these of existing state-of-the-art networks.

3.2 MAGNet: A Convolutional Neural Network with Global Attention Modules

In this subsection, we propose an architecture, MAGNet, a U-Net based network, that incorporates a multi-scale attention (MSA), a high-level feature improvement (HFIM), and a global attention network (GAN) to enhance the quality of biomedical image segmentation. The overall architecture of MAGNet is shown in Figure 3.1. The proposed MAGNet network includes modules for a high-level feature improvement (HFIM) [42] shown in Figure 3.2, a multi-scale attention (MSA) shown in Figure 3.3, and a global attention network (GAN) [42] shown in Figure 3.4. The features from the preprocessed images are first extracted using convolutional layers and then fed to be MSA module which gathers context, that is crucial for understanding complex images. The context thus obtained are used by HFIM module to improve the high-level features and capture even the smallest details. The GAN module provides comprehensive details of the image context, which is essential

for accurate segmentation of specific areas in complex biomedical images. Through the fusion of HFIM and GAN modules, MAGNet offers a comprehensive solution for biomedical image segmentation. It ensures accurate feature extraction, awareness of multi-scale context, improvement of complex structures, and a holistic understanding of the image. The network addresses a wide range of challenges and has the potential to outperform existing networks in various biomedical image segmentation tasks.

3.2.1 Architecture of MAGNet

MAGNet follows an encoder-decoder architecture with attention and feature enhancement modules to improve segmentation accuracy across different medical image modalities. The scheme for carrying out the segmentation of medical images by MAGNet is given in Figure 3.1 and details of the MAGNet network architecture are shown in Table 3.1.

The input image of size 256×256×3 is progressively downsampled through four encoding blocks. Each block consists of two 3×3 convolutional layers, batch normalization, and ReLU activation, followed by 2×2 max pooling. These layers extract increasingly abstract features while reducing spatial resolution. The output feature sizes evolve from 128×128×16 to 16×16×64.

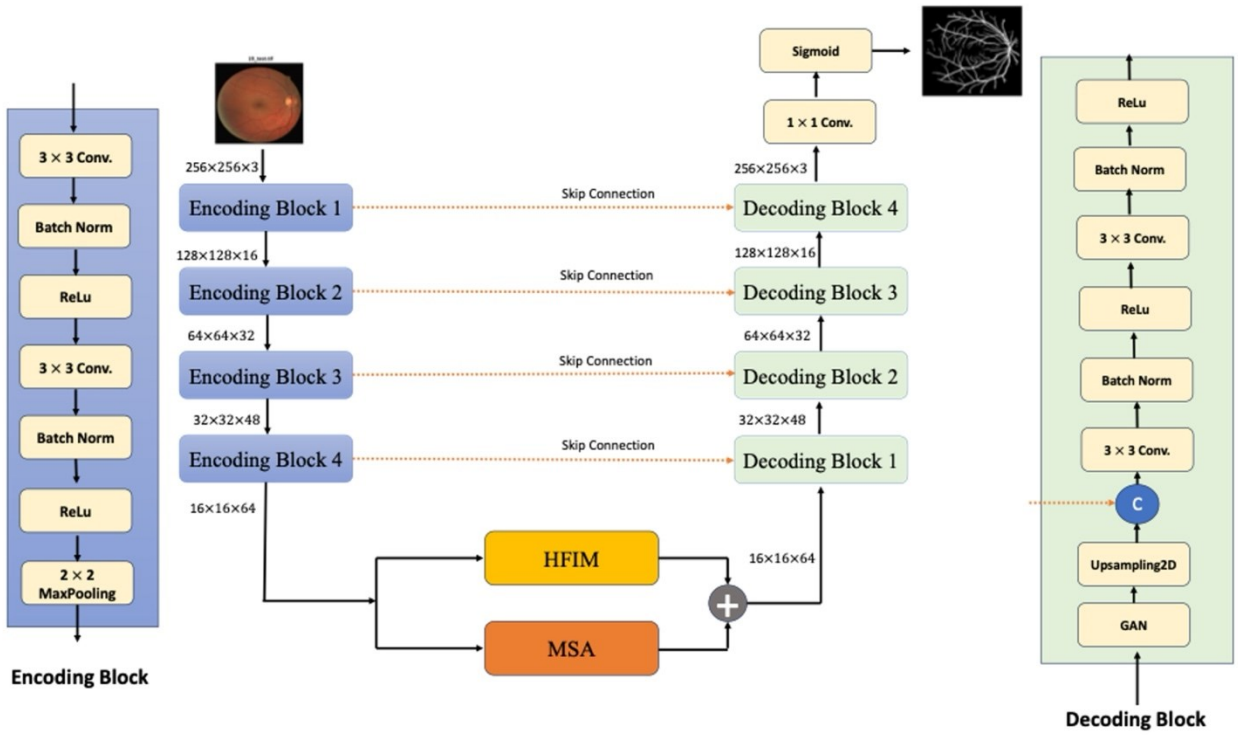


Figure 3.1: Architecture of MAGNet

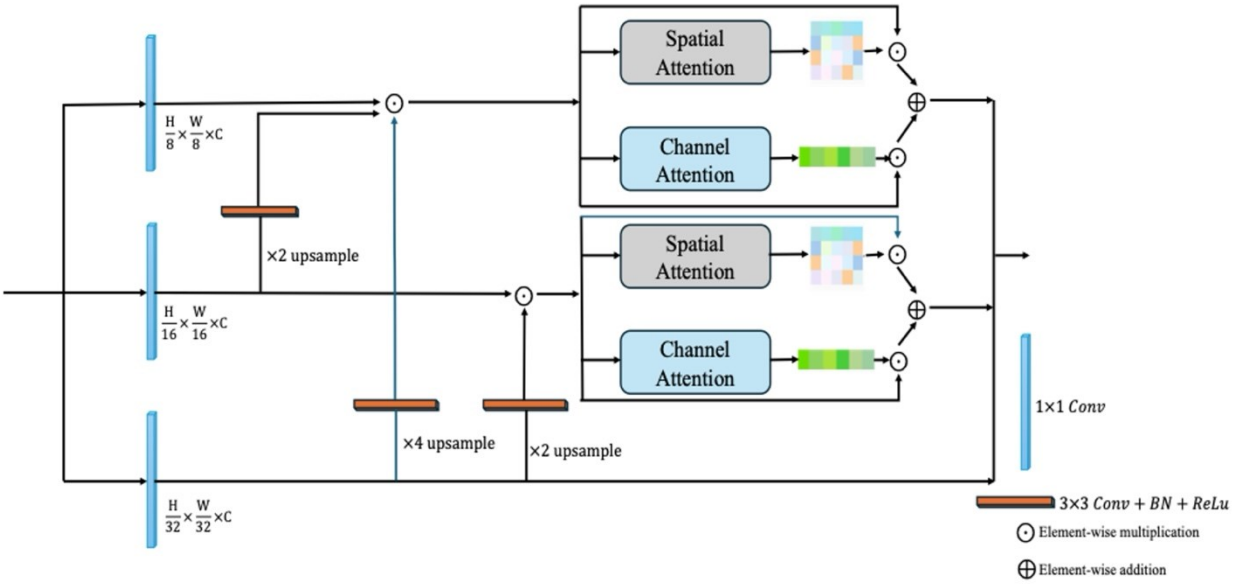


Figure 3.2: High-level feature improvement module (HFIM)

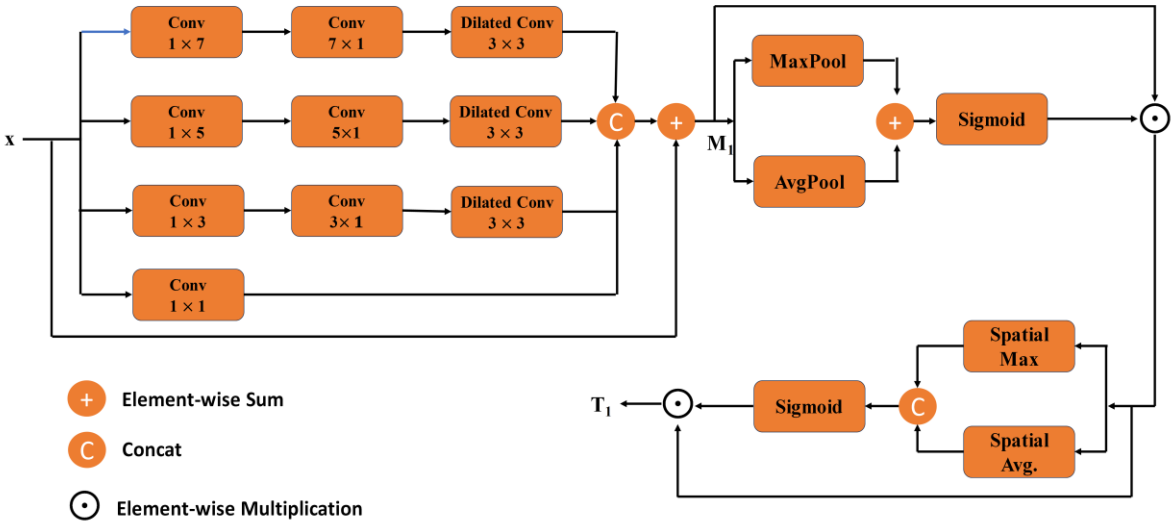


Figure 3.3: Architecture of MSA module

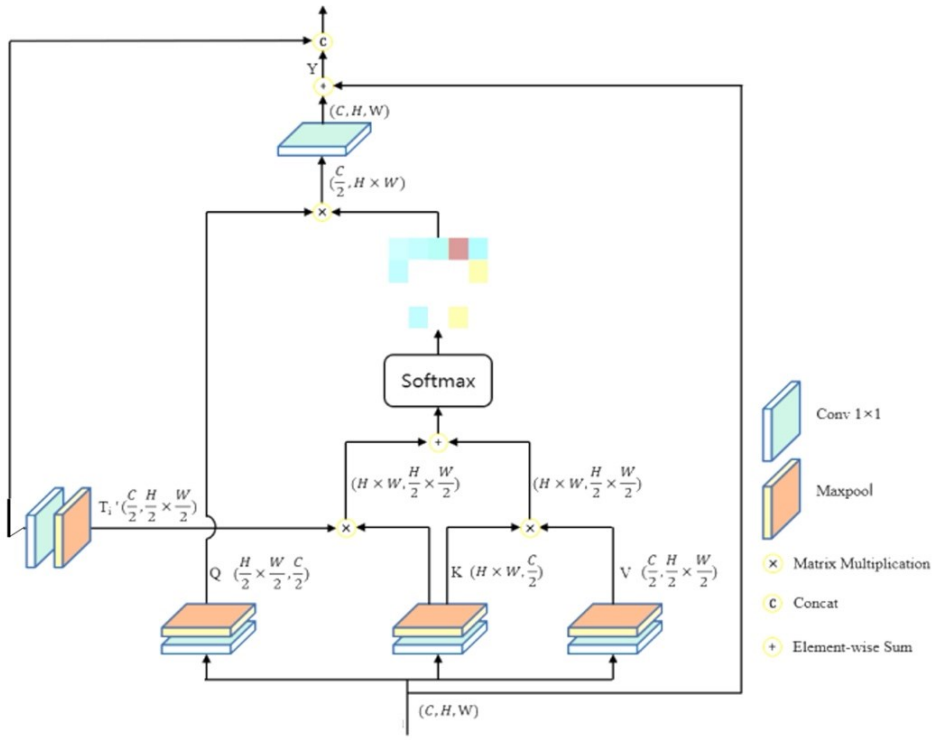


Figure 3.4: Global attention network (GAN)

Table 3.1: Details of the MAGNet network architecture

Block	**Layer**	**Number of input channels**	**Number of output channels**	**Input Size**	**Output Size**
Encoding Block 1	3×3 Conv +BN+ReLu	3	16	256×256	256×256
	3×3 Conv +BN+ReLu	16	16	256×256	256×256
	2×2 Maxpooling	16	16	256×256	128×128
Encoding Block 2	3×3 Conv +BN+ReLu	16	32	128×128	128×128
	3×3 Conv +BN+ReLu	32	32	128×128	128×128
	2×2 Maxpooling	32	32	128×128	64×64
Encoding Block 3	3×3 Conv +BN+ReLu	32	48	64×64	64×64
	3×3 Conv +BN+ReLu	48	48	64×64	64×64
	2×2 Maxpooling	48	48	64×64	32×32
Encoding Block 4	3×3 Conv +BN+ReLu	48	64	32×32	32×32
	3×3 Conv +BN+ReLu	64	64	32×32	32×32
	2×2 Maxpooling	64	64	32×32	16×16
HFIM		64	64	16×16	16×16
MSA		64	64	16×16	16×16
Decoding Block 1	GAN+ Upsample 2D + Concat	64	48	16×16	32×32
	3×3 Conv +BN+ReLu	48	48	32×32	32×32
	3×3 Conv +BN+ReLu	48	48	32×32	32×32

Decoding Block 2	GAN+ Upsample 2D + Concat	48	32	32×32	64×64
	3×3 Conv +BN+ReLu	32	32	64×64	64×64
	3×3 Conv +BN+ReLu	32	32	64×64	64×64
Decoding Block 3	GAN+ Upsample 2D + Concat	32	16	64×64	128×128
	3×3 Conv +BN+ReLu	16	16	128×128	128×128
	3×3 Conv +BN+ReLu	16	16	128×128	128×128
Decoding Block 4	GAN+ Upsample 2D + Concat	16	3	128×128	256×256
	3×3 Conv +BN+ReLu	3	3	256×256	256×256
	3×3 Conv +BN+ReLu	3	3	256×256	256×256
Final Block	3×3 Conv + Sigmoid	3	1	256×256	256×256

The output from the last encoding block is processed by two enhancement modules, HFIM and MSA. In this network, MSA and HFIM modules [41] sequentially process the output of the last convolutional layer. Next, the decoding block reconstructs the spatial dimensions using GAN, up-sampling, and concatenation operations, progressively taking inputs from the previous step and skip connections from the encoder. The last step consists of the final convolution and sigmoid activation to produce the segmented image.

Figure 3.2 illustrates the architecture of the High-Level Feature Improvement Module (HFIM) [42], developed to enhance multi-scale semantic features from deeper encoder layers. The module aggregates feature maps from three different resolutions, each with 64 channels. These features are

first upsampled to a common resolution and aligned through 3×3 convolutional blocks, followed by batch normalization and ReLU activation. To refine the aggregated features, each pathway incorporates channel attention and spatial attention modules, which adaptively reweights the feature responses. The outputs of both attention modules are fused via element-wise multiplication and summed to enhance important semantic information. This module significantly improves feature discrimination by leveraging contextual dependencies across both channel and spatial dimensions, enabling more accurate and robust segmentation in complex medical images.

In addition, we develop a MSA module that focuses on enhancing shallow features. The detailed architecture of MSA is depicted in Figure 3.3. This module is obtained to capture features across multiple scales through the incorporation of four convolutional layers, each characterized by progressively increasing kernel sizes and the application of inflated convolution techniques. These layers employ a variety of convolutional kernel sizes ($k \in \{1, 3, 5, 7\}$), each providing a distinct range of receptive fields. The design employs both $k \times 1$ and $1 \times k$ convolutions to achieve a substantial reduction in parameter count while minimally affecting the performance. Additionally, the use of dilated convolution in certain branches further extends the receptive field, enhancing the ability of network to distinguish features. The outputs from the multi-scale processing are merged along the channel dimension and integrated with x through pixel-wise addition, serving as residual connections. The subsequent attention mechanism then filters out background disturbances from these processed feature maps, highlighting essential details. It comprises both channel and spatial attention components, referred to as $Atten_c(\bullet)$ and $Atten_s(\bullet)$, respectively, as represented by

$$T_1 = Atten_s(Atten_c(M_1)) \tag{3.1}$$

The output of the multi-scale modules feature map is M_1. The channel attention mechanism is composed of an average-pooling layer ($AvgPool(\bullet)$), a max-pooling layer ($MaxPool(\bullet)$) and a Sigmoid activation function ($Sigmoid(\bullet)$), with its operational formula detailed by

$$Atten_c(x) = x \odot Sigmoid(AvgPool(x) + MaxPool(x)) \tag{3.2}$$

wherein input feature map is denoted by x. Following the application of channel attention, the weighted feature map is directed towards the spatial attention module. This sequence of operations is concisely represented by

$$Atten_s(x) = x \odot Sigmoid(Concat\big(AvgSpatial(x), MaxSpatial(x)\big)) \tag{3.3}$$

wherein $Concat(\bullet)$ denotes the process of concatenating features along the channel dimensions. $MaxSpatial(\bullet)$ refers to the operation of obtaining the maximum value across spatial dimensions, while $AvgSpatial(\bullet)$ calculates the average value across spatial dimensions.

Figure 3.4 illustrates the architecture of the Global Attention Network (GAN) [42], which enhances feature representations by modeling long-range dependencies across spatial and channel dimensions. The input feature map of size C, H, W is first processed through spatial downsampling using max-pooling operation and channel reduction through 1×1 convolution to generate the Query (Q), Key (K), and Value (V) representations. These tensors are reshaped and multiplied to compute a global attention map using a softmax operation, capturing contextual correlations between all spatial positions. Output of the previous operation is then elementwise added to the original feature map, keeping original information while incorporating global features. The modular design ensures efficient computation by operating on reduced spatial and channel dimensions. This attention mechanism allows the network to adaptively emphasize informative regions while suppressing less

relevant features, improving segmentation performance across complex medical imaging modalities.

The decoder in the proposed architecture in Figure 3.1 progressively restores spatial resolution using upsampling and convolutional operations, supported by skip connections from the encoder to retain fine-grained details. Each decoding block includes convolutional layers, batch normalization, ReLU activation, and an attention-guided refinement through the Global Attention Network (GAN) [23]. The final output is generated using a 1×1 convolution followed by a sigmoid activation to produce the binary segmentation map, accurately isolating the target regions in the medical image. The decoding block reconstructs the segmentation map by progressively upsampling the features to match the original image resolution. It uses transposed convolutions through Upsampling2D, followed by convolutional layers with batch normalization and ReLU activation. Skip connections from corresponding encoder blocks are concatenated with each decoding block to retain fine-grained spatial details. Finally, a 1×1 convolution followed by a sigmoid activation generates the binary segmentation map.

3.2.2. Description of Datasets

Three different modality datasets [43], [44], and [45] are used for experimental purposes.

The first dataset is the CT scans of the lung [43], referred to as dataset 1. This dataset consists of 829 CT scan slices of which 373 have been identified as positive, and the dataset has the ground truth masks of these positively identified slices. In our training stage, we use 85% of the CT scans and the remaining for testing. This dataset has been designed by [43] for the purpose of segmenting lung infection of COVID-19 patients.

The second dataset is the CVC-ClinicDB [44] dataset and referred to as the dataset 2 contains 612 polyp images along with their ground truth segmentation masks. For our experiment, we use 70% of this data for training our proposed model and the rest for testing. This dataset has been designed by [44] for the purpose of segmenting polyp from colonoscopy images.

The final dataset is the DRIVE dataset [45], referred to as dataset 3, contains 40 retinal images, split equally for the purpose of training and testing, along with the ground truth segmentation of their retinal trees. This dataset has been designed by [45] for the purpose of segmenting retinal blood for diagnosing hypertension or diabetic retinopathy.

3.2.3. Data Preprocessing and Training

In this subsection, we outline the procedure for preprocessing the images in all the three datasets. First, we resize each of the images in these datasets to the dimension of 224×224×3 pixels. For the CT scan images, various data augmentation techniques such as image scaling, rotation, addition of Gaussian noise, and cropping are employed to increase the number of images in the training set. For the DRIVE dataset, we use optical distortion, horizontal and vertical flipping, grid distortion, and elastic transform for augmentation. But, for the CVC-ClinicDB dataset, no augmentation is applied.

Our proposed network is trained separately on each of the three types of biomedical image datasets, resulting in three different trained models of MAGNet. We use binary cross entropy as the loss function and the AdamW optimizer with a learning rate of 1e-5, and a batch size of 1 during the training of the network for all the three datasets.

For segmenting the different modalities of data, make use of several Python library packages, such as Matplotlib, Sklearn, Numpy, Tensorflow, Keras, and many more. Cloud-based platform Google Collaboratory, with its GPU, is used for the experiments.

3.2.4. Results and Comparisons

Results of the Proposed Scheme and Ablation Study

Table 3.2 gives the results of the ablation study on the effectiveness of the global attention module on the performance of our proposed network in terms of intersection of union (IoU), dice coefficient (DSC), and accuracy (Acc) across a range of biomedical images. This finding highlights the importance of the attention mechanism in improving the performance of the proposed network.

Table 3.2: The performance results of the proposed MAGNet network with and without global attention module

Network	**Dataset 1**			**Dataset 2**			**Dataset 3**		
	IoU	**DSC**	**Acc**	**IoU**	**DSC**	**Acc**	**IoU**	**DSC**	**Acc**
MAGNet (without global attention module)	0.68	0.65	0.58	0.70	0.75	0.63	0.53	0.56	0.49
MAGNet	0.90	0.85	0.93	0.92	0.96	0.94	0.81	0.72	0.84

Comparison with existing state-of-the-art networks

Table 3.3 shows the performance results for the proposed MAGNet network in terms of IoU, DSC and accuracy, as well as in terms of the number of parameters for dataset 1 (CT scans). In this table, we also provide the corresponding results for Inf-Net [7], Gated UNet [8], DeepLab3+ [9], and

DenseUNet [10]. It is seen that the proposed MAGNet network provides the highest values for all the three metrics, significantly outperforming the other four networks.

Table 3.4 shows the performance results for proposed network for dataset 2 (CVC-ClinicDB). We also provide the corresponding results for ColonSegNet [11], PraNet [12], and DeepLab3+ [13]. It is seen from this table that the proposed network again provides the highest values for all the three metrics.

Table 3.5 shows the performance results for proposed network for dataset 3 (DRIVE). We also provide the corresponding results for CS-Net [14], SGL [15], RV-GAN [16], Attention UNet [17], and FR-UNet [18]. It is seen from this table that the proposed network again provides the highest values for all the three metrics.

It is seen from these three tables that, even though the proposed network utilizes the largest number of parameters, its performance is superior to that of all the other networks used for comparison, irrespective of the datasets.

Table 3.3: Results and comparisons of proposed network with existing networks for dataset 1 (CT scans)

Network	**No. of parameters**	**IoU**	**DSC**	**Acc**
Inf-Net [7]	33.12 M	0.78	0.68	0.69
Gated UNet [8]	175.09 K	0.72	0.63	0.65
DeepLab3+ [9]	43.9 M	0.87	0.78	0.73
DenseUNet [10]	45.08 M	0.65	0.52	0.84
MAGNet (Proposed)	55.5 M	0.90	0.85	0.93

Table 3.4: Results and comparisons of proposed network with existing networks for dataset 2 (CVC-ClinicDB)

Network	No. of parameters	IoU	DSC	Acc
ColonSegNet [11]	5.01 M	0.83	0.89	0.81
PraNet [12]	32.55 M	0.88	0.92	0.90
DeepLab3+ [13]	39.76 M	0.89	0.93	0.91
MAGNet (Proposed)	55.5 M	0.92	0.96	0.94

Table 3.5: Results and comparisons of proposed network with existing networks for dataset 3 (DRIVE)

Network	No. of parameters	IoU	DSC	Acc
CS-Net [14]	8.40 M	0.70	0.63	0.75
SGL [15]	15.53 M	-	0.67	0.69
RV-GAN [16]	14.81M	-	0.69	0.65
Attention UNet [17]	8.73 M	0.67	0.59	0.65
FR-UNet [18]	5.72 M	0.71	0.70	0.72
MAGNet (Proposed)	55.5 M	0.81	0.75	0.84

Visual Analysis

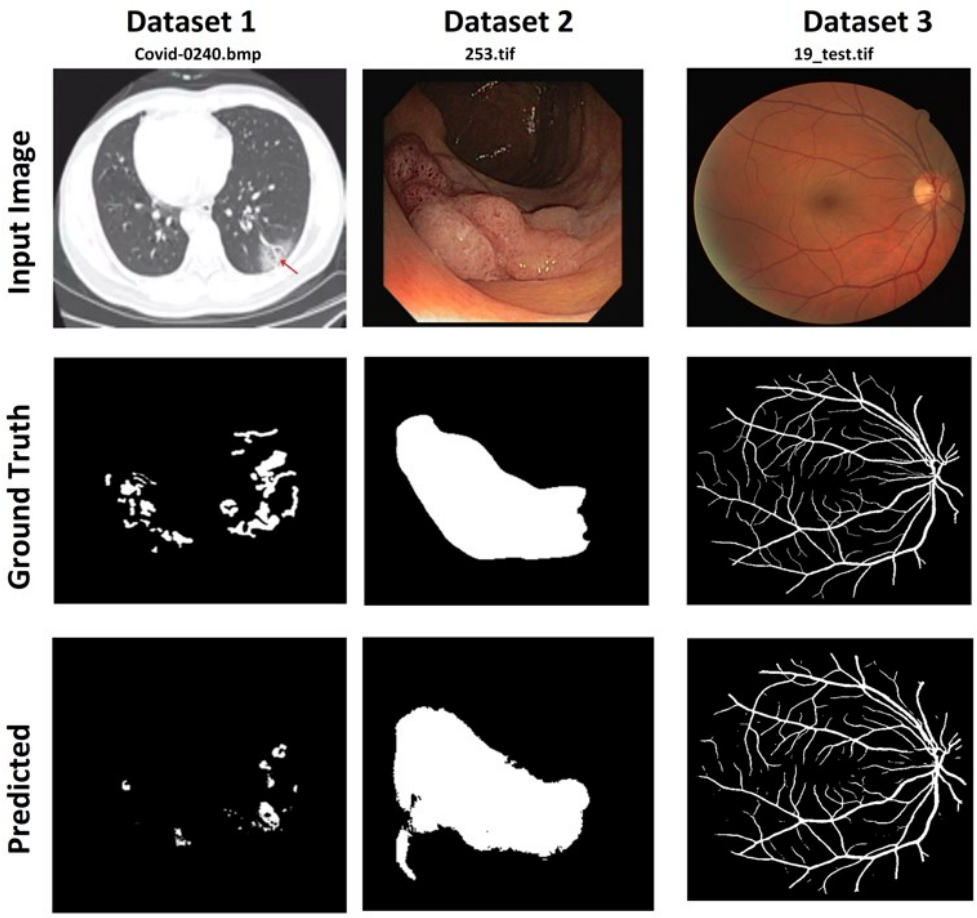


Figure 3.5: Visual illustration of the segmentation performance of the proposed MAGNet network

Figure 3.5 provides a visual illustration of the segmentation performance of the proposed network. For this purpose, we have selected one typical image from each of the three datasets shown in the first raw of the figure. The second and the third rows of this figure illustrate the corresponding ground truth segmentation results and the corresponding results obtained by the proposed network. By comparing the results in second and third rows, it is seen that the proposed network is able to satisfactorily segment the regions of interest.

3.3 MedSegNet: A Convolutional Neural Network with Dual Self-Attention

Despite the good performance of MAGNet in terms of IoU, DSC and Acc, the number of parameters used is rather high. We propose in this section a lighter weight network, MedSegNet, whose performance is about the same as that of MAGNet. This network employs a dual self-attention module instead of HFIM and GAN to significantly reduce the computational complexity. This design ensures that the network is not only effective, but also accessible for real-time applications and can be deployed in environments with limited computational resources. Incorporating the residual blocks within the U-shaped architecture improves the ability of network to learn from a wide range of features by facilitating deeper network architectures. This integration reduces the training and testing errors, ensuring that the performance of network improves with depth, making it robust to various imaging modalities and conditions. Dual self-attention mechanism of MedSegNet employs both spatial and channel-wise attention, allowing the network to dynamically focus on salient features while suppressing irrelevant information. This attention to detail is particularly crucial in medical imaging, where the distinction between pathological and non-pathological regions can be subtle. Our network further integrates a multi-scale attention (MSA) mechanism, enabling it to effectively capture features at various scales and resolutions. This capability is critical for accurately segmenting medical images, where important details may be present at different scales, from local to global.

3.3.1 Architecture of MedSegNet

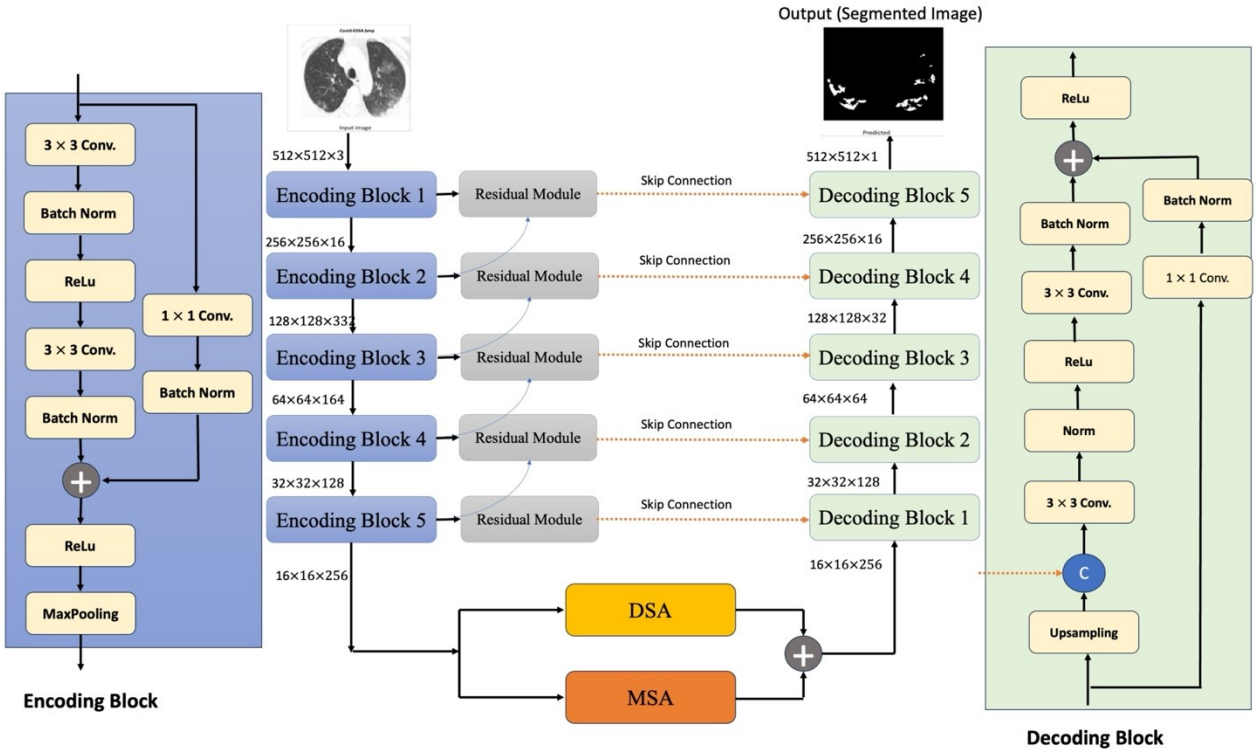


Figure 3.6: Proposed MedSegNet network

Figure 3.6 illustrates the overall structure of the proposed MedSegNet network. Using a U-shaped framework enhanced with skip connections forms the base of our network structure, incorporating five convolutional blocks for each of the encoding and decoding stages. Each encoding block is structured around a primary encoding pathway and an additional residual pathway. Sequential 3×3 convolutions process the incoming feature map within the main pathway, with the initial convolution followed by batch normalization and ReLU activation, while the subsequent convolution is succeeded by normalization alone. To reduce the issue of network errors, a residual pathway is integrated within the encoding block. The decoding block mirrors the encoding

structure, differing only in its inclusion of bilinear interpolation for up-sampling purposes towards the end.

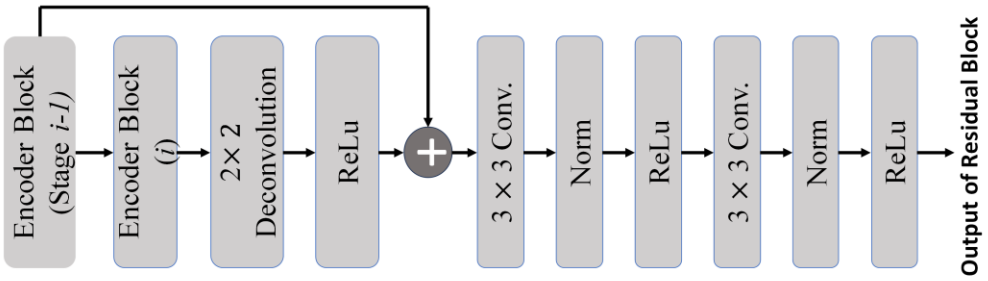


Figure 3.7: Structure of the residual module

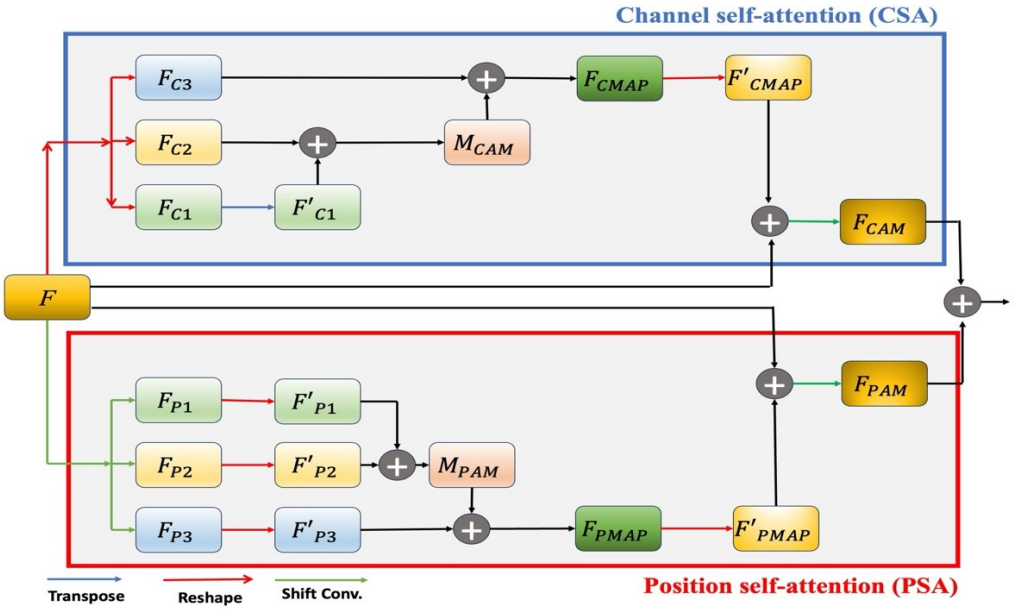


Figure 3.8: Architecture of DSA module

The detailed structure of the residual module is presented in Figure 3.7. A residual block-enhanced skip connection is implemented between encoding and decoding blocks at corresponding stages. This process involves up-sampling the reduced feature map from the subsequent stage by 2×2

deconvolution and ReLU activation, before subtracting this from the output of the current stage's encoding block to recapture details lost during down-sampling. To further highlight residual feature representation, a convolutional block comprising dual 3×3 convolution layers is employed in Figure 3.7. The implementation of a DSA module [47] and an MSA module [37], [48] helps to establish a wide range of dependencies. The DSA module, shown in Figure 3.8, independently constructs long-range dependencies across channel and spatial dimensions. The synthesized features from these modules are then fed into the decoding block, resulting in the segmentation of images through a sequential application of decoding blocks. In the task of segmentation, shallow features play a critical role in enhancing the segmentation accuracy. These features contain a lot of information concerning textures, edges, and other critical attributes. Simultaneously, the significance of multi-scale features cannot be overstated, given the notable variation in images belonging to different modality. Effectively, the model must precisely maintain the complicated details of smaller affected regions against the backdrop of complex background noise, while also ensuring the comprehensive representation and precise outlining of the larger boundaries of affected areas.

Figure 3.8 illustrates the architecture of the dual self-attention (DSA) module [49] used in the architecture of MedSegNet. This module is designed to enhance feature representation by capturing long-range dependencies across both channel and spatial dimensions. The input to the module is a feature map denoted as $F \epsilon R^{C \times H \times W}$, where C, H, and W represent the number of channels, height, and width, respectively. This input is simultaneously fed into the channel self-attention (CSA) and the position self-attention (PSA) blocks. In the CSA block whose boundary is shown in blue, the input feature F is sequentially processed by two fully connected layers, producing the intermediate features F_{c1} and F_{c2}, respectively. These are then element-wise added and passed through another

fully connected layer F_{c3}, resulting in a channel attention map M_{CAM}. This attention map is multiplied elementwise with the original input F to obtain the channel-refined feature map F_{FMAP}. The channel attention-enhanced output F_{CAM} is then generated by adding F_{FMAP} to the original input F, incorporating both the local and global channel contexts. In the PSA block whose boundary is shown in red, the input feature F is transformed through three convolutional layers F_{p1}, F_{p2}, and F_{p3}, generating three projections F'_{p1}, F'_{p2} and F'_{p3}, respectively. These projections are reshaped and combined through a matrix multiplication followed by a softmax activation to yield the position attention map M_{PAM}. This map is then used to reweight the input features, resulting in the position-refined feature map F_{PMAP}. The final position-enhanced feature F_{PAM} is obtained by adding F_{PMAP} to the original input F. Finally, the outputs from both of the CSA and PSA blocks, namely, F_{CAM} and F_{PAM}, are combined through element-wise addition to produce the final output of the DSA module. This DSA module enables MedSegNet to effectively capture inter-channel dependencies as well as spatial relationships, thereby improving the representation of the feature maps for segmentation. The decoder of the proposed architecture, as depicted in the given Figure 3.6, is responsible for progressively reconstructing the segmented output from the compact and abstract feature maps. After the aggregated features from the DSA and MSA modules are combined, the decoding process begins with decoding block 1, which receives input of size 16×16×256. Each decoding block consists of a sequence of convolutional operations, batch normalization, and ReLU activations, followed by an upsampling operation to increase the spatial resolution. At each decoding block, the upsampled output is concatenated with the corresponding skip connection feature map from the encoding block, enabling the network to retain high-resolution spatial details lost during downsampling. Such decoding continues through decoding blocks 3, 4, and 5, progressively reconstructing the segmentation map to the output image size of 512×512×1. The

final layer applies a 1×1 convolution to map the multi-channel feature representation to a segmentation output. Overall, the decoder effectively leverages both hierarchical context and fine-grained spatial features to produce accurate and well-localized segmentation results.

3.3.2. Data Preprocessing and Training

The datasets employed for training MedSegNet are the same as the ones used for training MAGNet, the detailed description of which are provided in Subsection 3.2.2. We first resize all the images of the three datasets to 512×512×3 pixels. We then apply various augmentation techniques to each of the datasets to enrich the data and prevent the model from overfitting. For the CT scan images, we use rotation, scaling, cropping, and adding Gaussian noise for augmenting the dataset. For the DRIVE dataset, we implement horizontal and vertical flipping, elastic transformations [50], grid distortion [51], and optical distortion [51] for augmenting the dataset. However, we do not need any augmentation to the CVC-ClinicDB dataset. For training the model, we use the Adam optimizer, a batch size of 8, and binary cross-entropy as the loss function during training, utilizing 85% of the samples for training.

Google Collaboratory is used for the implementation. This is a Jupyter notebook-based cloud service for sharing information and implementing deep learning (DL). It handles DL workloads that are GPU-bound and provide completely optimized runtimes. We use the library packages keras, Tensorflow, Sklearn, Matplotlib, and Numpy for the implementation purposes.

3.3.3. Results and Comparisons

Results of the Proposed Scheme and Ablation Study

Table 3.6 depicts the results of ablation study on the performance of MedSegNet by adding progressively to the base model, the residual module, the DSA module and MSA module. It is seen from the table that the performance of the proposed MedSegNet network progressively improves by the inclusion of the three modules.

Table 3.6: Impact of different modules on segmentation performance

Model	CT scan		DRIVE		CVC-ClinicDB	
	DSC	**IoU**	**DSC**	**IoU**	**DSC**	**IoU**
Base Model	0.51	0.63	0.54	0.57	0.89	0.84
Base Model + Residual Networks	0.53	0.66	0.55	0.59	0.91	0.90
Base Model + Residual Modules + DSA	0.77	0.79	0.70	0.73	0.92	0.91
Base Model + Residual Modules + DSA+MSA (Proposed)	0.83	0.89	0.73	0.78	0.94	0.90

Comparison with Existing State-of-the-art Networks

In this section, we discuss the performance of our MedSegNet network in segmenting the objects of interest, namely, lung infection using CT scan, retinal diseases using DRIVE, and polyp using

CVC-CLINICDB datasets. The segmentation performance results for the CT scan dataset using the MedSegNet as well as the state-of-the-art networks, DenseUNet [10], Gated UNet [8], Inf-Net [7], and DeepLab3+ [9] are given in Table 3.7. It is seen that the MedSegNet network outperforms that of the state-of-the-art networks using the lowest number of parameters.

The segmentation performance results for the DRIVE dataset using the MedSegNet as well as the state-of-the-art networks, Attention UNet [17], CS-Net [14], RV-GAN [16], SGL [15], and FR-UNet [18] are given in Table 3.8. It is seen that the MedSegNet network outperforms that of the state-of-the-art networks.

The segmentation performance results for the CVC-CLINICDB dataset using the MedSegNet as well as the state-of-the-art networks, ColonSegNet [11], PraNet [12], and DeepLab3+ [13] are given in Table 3.9. It is seen that the MedSegNet network outperforms that of the state-of-the-art networks.

Table 3.7: Results and comparisons with state-of-the-art networks using the CT scan dataset

Model	**No. of parameters**	**DSC**	**IoU**	**Acc**
DenseUNet [10]	45.08 M	0.52	0.65	0.85
Gated UNet [8]	175.09 K	0.63	0.72	0.65
Inf-Net [7]	33.12 M	0.68	0.78	0.69
DeepLab3+ [9]	43.9 M	0.78	0.87	0.87
MAGNet (Proposed)	55.5 M	0.85	0.90	0.93
MedSegNet (Proposed)	16 M	0.83	0.89	0.95

It is also seen from Tables 3.7, 3.8, 3.9 that the segmentation performance in terms of DSC, IoU and Acc of both our proposed networks, MAGNet and MedSegNet, are about the same. However, MedSegNet is a significantly lower weight network, requiring only 16 million parameters compared to MAGNet's 55.5 million parameters, those providing a computationally efficient architecture.

Table 3.8: Results and comparisons with state-of-the-art networks using the DRIVE dataset

Model	No. of parameters	DSC	IoU	Acc
Attention UNet [17]	8.73 M	0.59	0.67	0.65
CS-Net [14]	8.40 M	0.63	0.70	0.75
RV-GAN [16]	14.81M	0.69	-	0.65
SGL [15]	15.53 M	0.67	-	0.69
FR-UNet [18]	5.72 M	0.70	0.71	0.72
MAGNet (Proposed)	55.5 M	0.75	0.81	0.84
MedSegNet (Proposed)	16 M	0.73	0.78	0.86

Table 3.9: Results and comparisons with state-of-the-arts networks using the CVC-ClinicDB dataset

Model	No. of parameters	DSC	IoU	Acc
ColonSegNet [11]	5.01 M	0.89	0.83	0.81
PraNet [12]	32.55 M	0.92	0.88	0.90
DeepLab3+ [13]	39.76 M	0.93	0.89	0.91
MAGNet (Proposed)	55.5 M	0.96	0.92	0.94
MedSegNet (Proposed)	16 M	0.94	0.90	0.96

Visual Analysis

Figure 3.9 provides a visual illustration of the segmentation performance of the proposed network. For this purpose, we have selected one typical image from each of the three datasets shown in the first column of the figure. The second and the third column of this figure illustrate the corresponding ground truth segmentation results and the corresponding results obtained by the proposed network. By comparing the results in second and third columns, it is seen that the proposed network is able to satisfactorily segment the regions of interest.

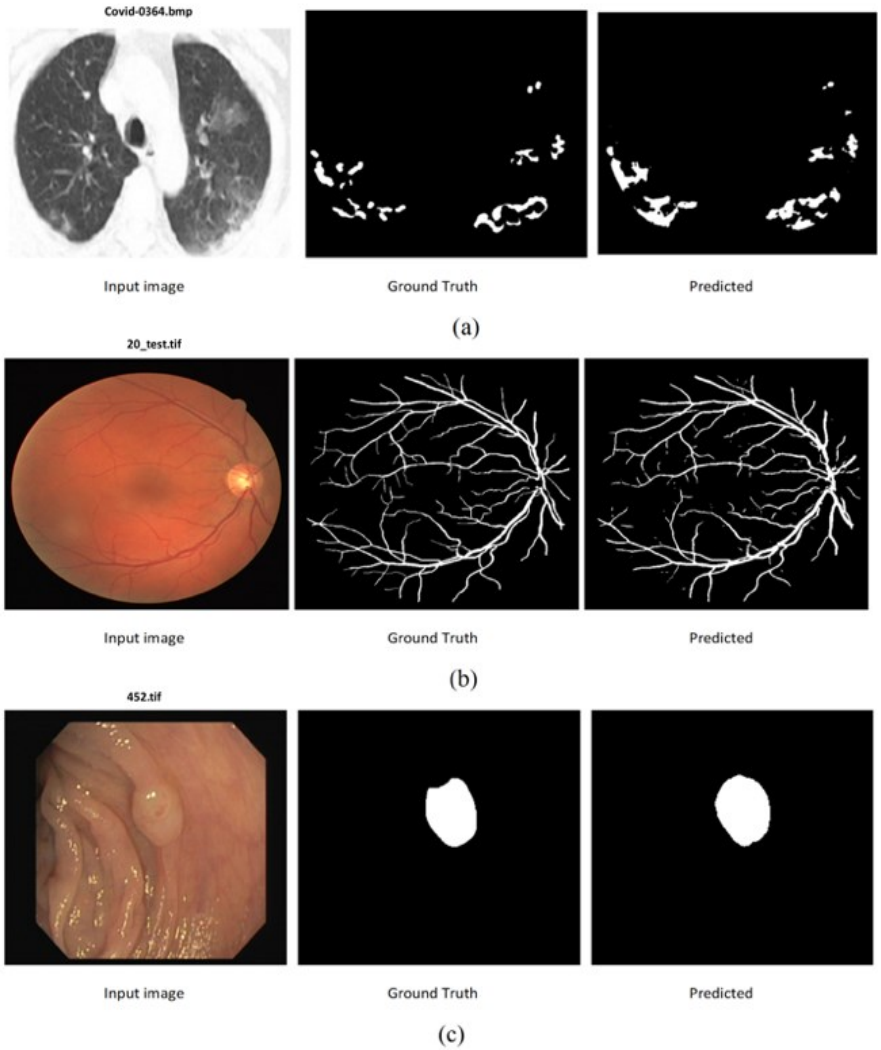


Figure 3.9: Qualitative results of (a) CT scan, (b) DRIVE, and (c) CVC-CLINICDB dataset for our proposed MedSegNet network

3.4 SSNet: A Semi-Supervised Convolutional Network

In this section, we propose a semi supervised network for segmentation of brain MR images to diagnose and monitor patients with neurological diseases. The purpose of implementing semi-supervised learning, rather than solely relying on supervised learning, is to effectively use a substantial amount of unlabeled data alongside a limited set of labeled data to enhance network performance. Noting that our brain MRI dataset contains significantly more unlabeled data than labeled data, semi-supervised learning presents an appropriate and effective strategy for handling such unlabeled data. This proposed scheme is an integrated multi-scale attention-enhanced semi supervised network called SSNet, segmenting the object of interest in MRIs of the brain. In our proposed network, multi-scale feature extraction is employed to extract features at multiple resolutions in order to capture both the global and local features while the attention mechanism enhances the focus of network on relevant structures, resulting in an improved segmentation performance.

3.4.1 Architecture of SSNet

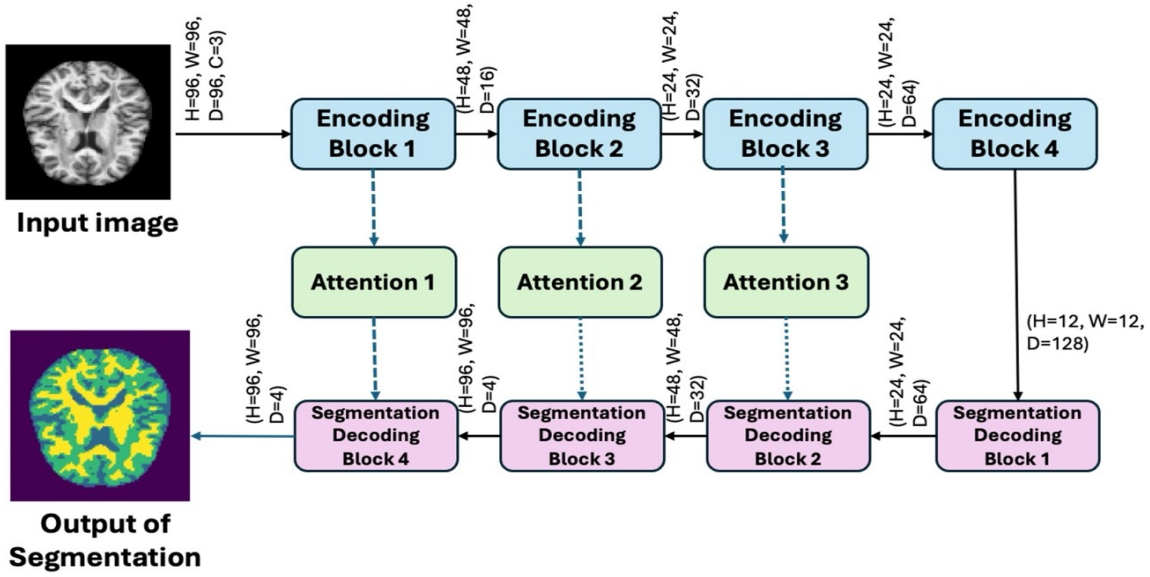
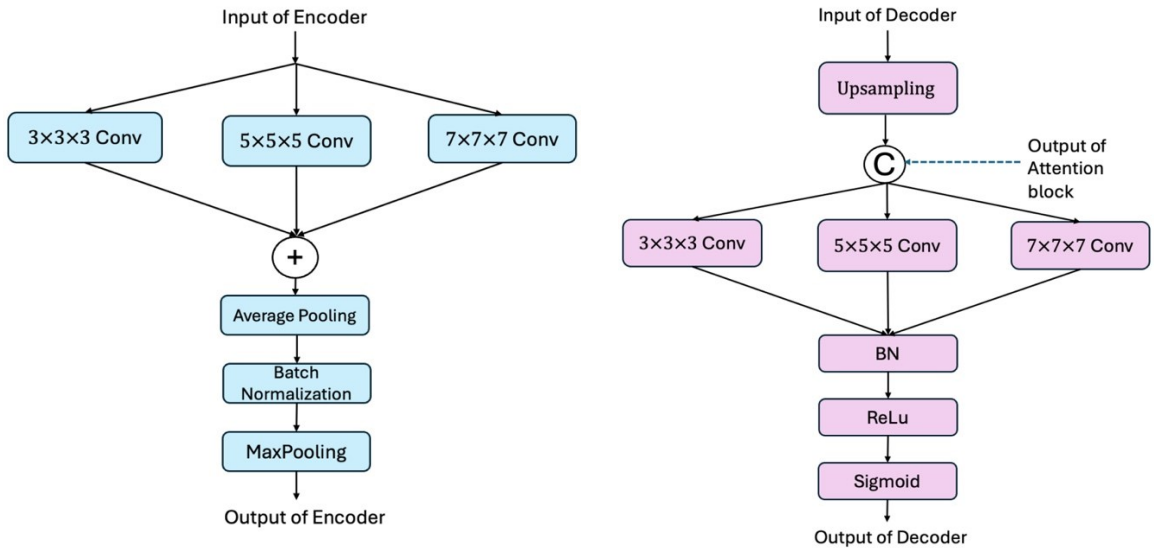


Figure 3.10: Proposed multi-scale attention-enhanced semi supervised network (SSNet)



(a) (b)

Figure 3.11: Architecture of (a) encoding and (b) decoding block

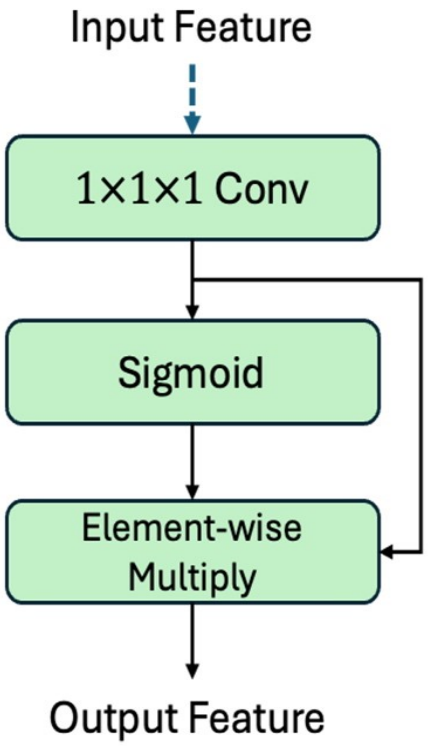


Figure 3.12: Attention block of SSNet network

The proposed multi-scale attention-enhanced semi supervised network, called SSNet, for segmentation of brain MRI images, where we have limited labeled data is shown in Figure 3.10. The input image of size 96×96×96 with 3 channels is fed to a series of four encoding blocks. Each encoding block progressively reduces the spatial dimensions while increasing the feature depth, allowing the network to capture increasingly abstract representations. Following each encoding block, an attention module is applied to enhance the feature maps by focusing on the most relevant regions, which helps in emphasizing the meaningful regions for segmentation. The outputs of the attention modules are then passed to corresponding segmentation decoding blocks. Each decoding block uses a different resolution scale, enabling the network to reconstruct the segmentation maps.

The SSNet architecture is specifically designed to utilize both labeled and unlabeled data effectively in a semi-supervised learning, making it well-suited for brain MRI segmentation with limited labeled data.

The details of an encoding block in Figure 3.10 shown in Figure 3.11 (a) consists of several stages, employing a multi-scale convolution. The input features are processed through parallel convolution layers with kernel sizes of 3×3×3, 5×5×5, and 7×7×7, capturing low level features to high level features in each encoding block. The outputs of these convolutions are averaged and passed through a batch normalization layer. This approach efficiently captures both the local and global features without significantly increasing the computational cost. The number of filters in the encoding blocks 1, 2, 3, and 4 are 3, 16, 32, and 64, respectively. Each encoding block of the network reduces the spatial resolution of the feature maps through a MaxPooling layer to gradually downsample the image as well as to capture hierarchical features. The downsampled feature maps are then passed through attention blocks to focus on the important regions. To improve the ability of the network to focus on key regions of the image, an attention block [23] shown in Figure 3.12 is added after each encoding block. The output of each encoding block is fed into the input to an attention block. The attention weights are obtained by a 1×1×1 convolution followed by a sigmoid activation block that normalizes the attention map values to the range [0, 1]. The output of sigmoid activation block is multiplied by the output of the 1×1×1 convolution, as a result of which the important features become more significant while the irrelevant ones disappear. Element-wise multiplication re-weights the input feature map by the learned attention map. The output of the attention block is an enhanced feature map emphasizing informative regions. The proposed SSNet network can be trained in an end-to-end manner, and the attention block helps the network to focus on different regions for better segmentation results. The decoding block shown in Figure 3.11(b) integrates the

multi-scale convolutional operations with an attention-enhanced upsampling pathway to refine the segmentation output. The output of encoding block 4 is fed as the input to decoding block 1, which uses 128 filters. This input is first upsampled to restore spatial resolution and then processed in parallel through three convolutional layers with kernel size 3×3×3, 5×5×5, and 7×7×7, to extract features at multiple receptive fields, allowing the network to capture both the local and global features. The outputs of these convolutional layers are passed through a batch normalization layer, followed by a ReLU activation for non-linearity, and finally a sigmoid activation to generate the final output of the decoding block, typically representing a probability map for segmentation. The output of the decoding block 1 is fed to a series of three decoding blocks, having 64, 32, and 4, respectively. These decoding blocks 2 to 4 employ a common architecture that facilitates the progressive reconstruction of spatial resolution in the segmentation process. Each of these decoding blocks receives the output from the previous decoding block and first applies an upsampling operation to increase the spatial dimensions. This upsampled feature map is then concatenated with the corresponding output of attention module, allowing for the recovery of fine-grained spatial details. The concatenated feature in each decoding block is passed through parallel multi-scale convolutional layers with kernel sizes 3×3×3, 5×5×5, and 7×7×7, enabling the network to capture both the local and the global features at varying receptive fields. The outputs from these convolutions are fused and passed through a batch normalization layer, followed by ReLU activation, and a Sigmoid activation to produce the output of each decoding block. Finally, the output of the decoding block 4 provides a segmented image or regions of interest in the original medical image.

3.4.2 Description of Dataset

The experimental dataset comprises a total of 466 3D brain MRI scans, including 416 real T1-weighted volumes from the publicly available OASIS-1 dataset [52] and 50 procedurally generated synthetic scans created using the MONAI framework [53], [54]. The synthetic images simulate realistic anatomical variations and are incorporated to enhance data diversity and reduce overfitting. By combining real and synthetic data, we ensure robust evaluation and improved performance for brain MRI image segmentation for the limited number of labeled data. This dataset has been designed by [52] for the purpose of segmenting brain tissue to diagnose abnormalities in the brain.

Table 3.10: Summary of the dataset that used for semi-supervised learning

Data Category	Count	% of Total (466)	Comments
Real images (OASIS-1)	416	-	T1-weighted 3D MRIs
Synthetic images	50	-	Procedurally generated using MONAI
Total images	466	100%	Combined dataset used for segmentation
Labeled data	186	40%	Have ground truth masks
Unlabeled data	280	60%	Without masks, used in semi-supervised training
Unlabeled data used for training	280	-	All unlabelled images are trained.
Labeled data used for training	148	80% of 186 total images	Supervised dice loss
Total number of images used for training	438	-	148 labeled (80% of total labeled images) + 280 unlabeled = 438 images for training
Unlabeled data used for testing	0	-	Not used directly in testing
Labeled data used for testing	38	20% of 186 total images	20% of labeled data (ground truth masks used for testing)

Total number of images used for testing	38	-	Data used for testing

3.4.3 Data Preprocessing and Training

As mentioned above, we have a total of 466 3D brain MRI images. Using trilinear interpolation [55], we reshape these 3D brain images to a uniform size of 96×96×96 voxels, ensuring that the network consistently receives inputs having the same dimensions. For the semi-supervised learning, 40% of the total number of images, (that is, 186 images), are considered as labeled data, each paired with a ground truth segmentation mask generated using MONAI [54]. The remaining 60% (that is, 280 images) are considered unlabeled, without any corresponding ground truth segmentation mask. For training the network, 80% of the labeled images (that is, 148 images) along with all the 280 unlabeled images are used. The remaining 20% of the labeled data (that is, 38 images) is used for testing our proposed network. The unlabeled data contributes only in training the network, but is not involved in the testing. The above information is given in Table 3.10.

We conduct the experiments in a cloud-based environment using NVIDIA Tesla V100 GPUs. We implement the network in PyTorch and the MONAI framework [54]. We train the proposed SSNet network with a batch size of 8 and a learning rate of 1e-3. We optimize the network using the Adam optimizer and the following proposed hybrid loss function which is a weighted combination of the dice loss [56] and the mean squared error (MSE) loss [57], and is given by:

$$\mathcal{L}_{hybrid} = \lambda_1 . \mathcal{L}_{dice}\left(S_{pred}, S_{true}\right) + \lambda_2 . \mathcal{L}_{MSE}\left(R_{pred}, R_{true}\right) \tag{3.4}$$

where $\mathcal{L}_{dice}$ is the dice loss between the predicted segmentation S_{pred} and the ground truth segmentation S_{true}, $\mathcal{L}_{MSE}$ is the MSE loss between the predicted displacement field R_{pred} and the

ground truth displacement field R_{true}, and λ_1 and λ_2 are scalar weights that control the relative importance of the segmentation. Assuming the both losses contribute equally to the overall loss, we set both λ_1 and λ_2 to unity.

3.4.4. Results and Comparison

Results of the Proposed Semi-Supervised Learning

Table 3.11 illustrates the results, in terms of DSC, IoU, precision and recall, of semi-supervised learning with different proportions of labeled and unlabeled data, for segmentation of brain MRI images. The results clearly demonstrate that semi-supervised learning allows the network to achieve superior performance with 60% unlabeled data, making it a promising approach for the segmentation of brain MRI images.

Table 3.11: Results of semi-supervised learning

Data Availability	**Dice Score (DSC) (%)**	**IoU (%)**	**Precision (%)**	**Recall (%)**	**No. of labeled and unlabeled images**
Semi-Supervised (40% labeled, 60% unlabeled)	88.36	79.83	88.75	88.35	186 labeled + 280 unlabeled=466
Semi-Supervised (20% labeled, 80% unlabeled)	82.75	71.96	82.90	84.28	93 labeled+373 unlabeled=466
Semi-Supervised (10% labeled, 90% unlabeled)	77.96	66.63	81.26	76.43	47 labeled+419 unlabeled=466

Comparison with Existing State-of-the-art Networks

The results of segmentation of the brain MRI dataset images using the SSNet as well as that of the state-of-the-art networks, namely, U-Net [21], DeepLab [58], and SegNet [59] are given in Table 3.12. It is seen that the SSNet network outperforms that of the state-of-the-art networks in terms of all the four matrices, DSC, IoU, precision and recall. It is also seen that the number of parameters used by the proposed network is smaller than that used U-Net and DeepLab and larger than that used by SegNet.

Table 3.12: Performance comparison of the proposed SSNet with that of the state-of-the-art networks

Network	**Number of parameters**	**Dice Score (DSC) (%)**	**IoU (%)**	**Precision (%)**	**Recall (%)**
U-Net [21]	31.04 M	85.26	78.92	84.03	86.57
DeepLab [58]	134.3M	86.88	80.19	85.31	88.03
SegNet [59]	1.6 M	84.51	76.87	83.29	85.05
Proposed SSNet Network	10.68 M	88.36	79.83	88.75	88.35

Visual Analysis

Figure 3.13 presents a visual illustration of the segmentation results obtained using the proposed SSNet network on brain MRI data. The first row displays the original slices of an image from three anatomical planes, the axial, coronal, and sagittal. The second row shows the corresponding ground truth segmentations, while the third row illustrates our predicted segmentation produced by the proposed SSNet network. A comparative analysis of the second and third rows indicates that our network can accurately isolate all views of anatomical structures of complex brain tissues.

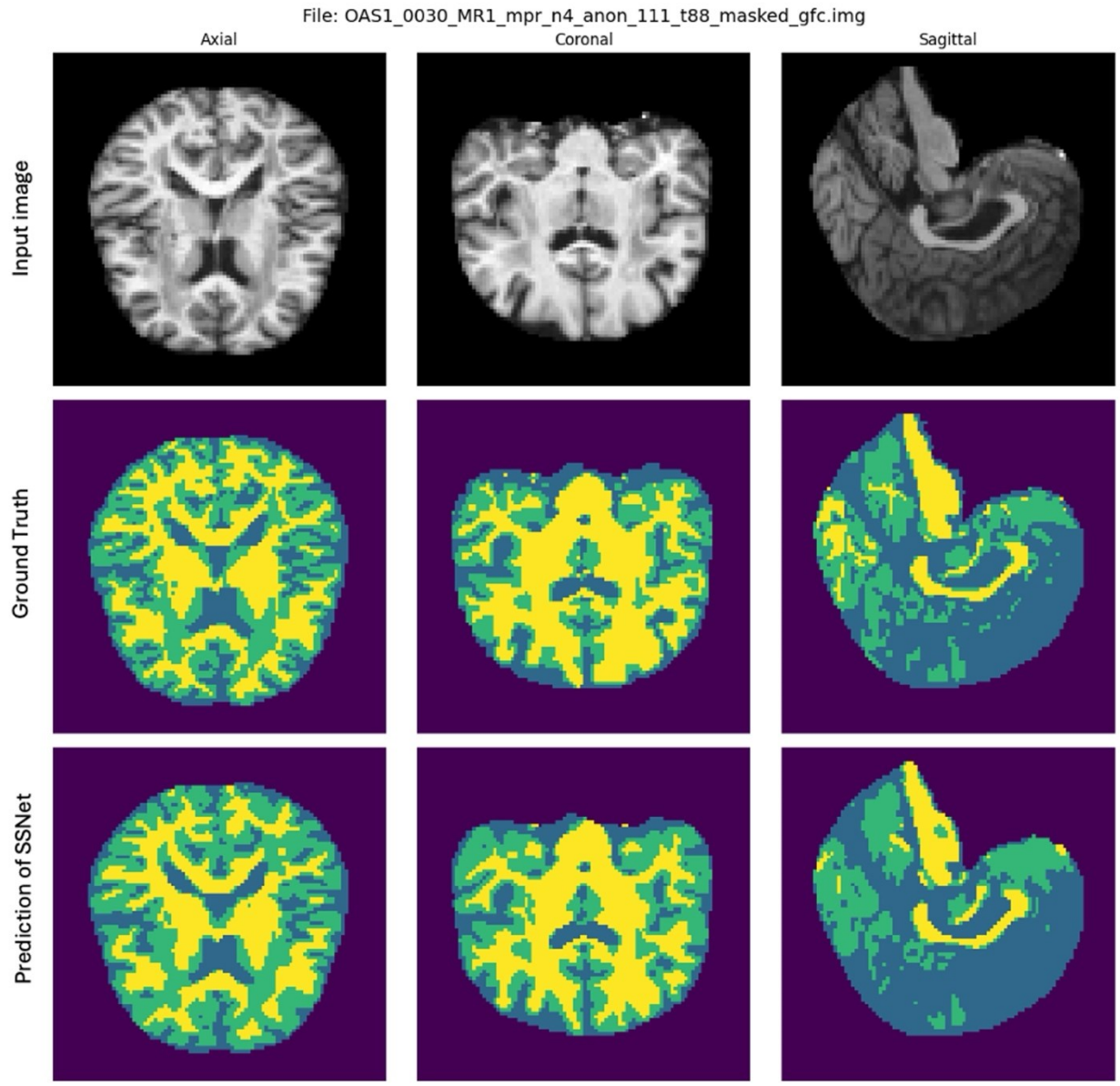


Figure 3.13: Visual representation of segmentation of brain MRI of SSNet network with respect to ground truth

3.5 FFNet: Convolutional Neural Network with Multipath Encoder

In this section, we implement a U-Net based multipath encoder with a feature fusion convolutional network that presents a novel scheme to segment breast cancer region from ultrasound images.

3.5.1 Architecture of FFNet Network

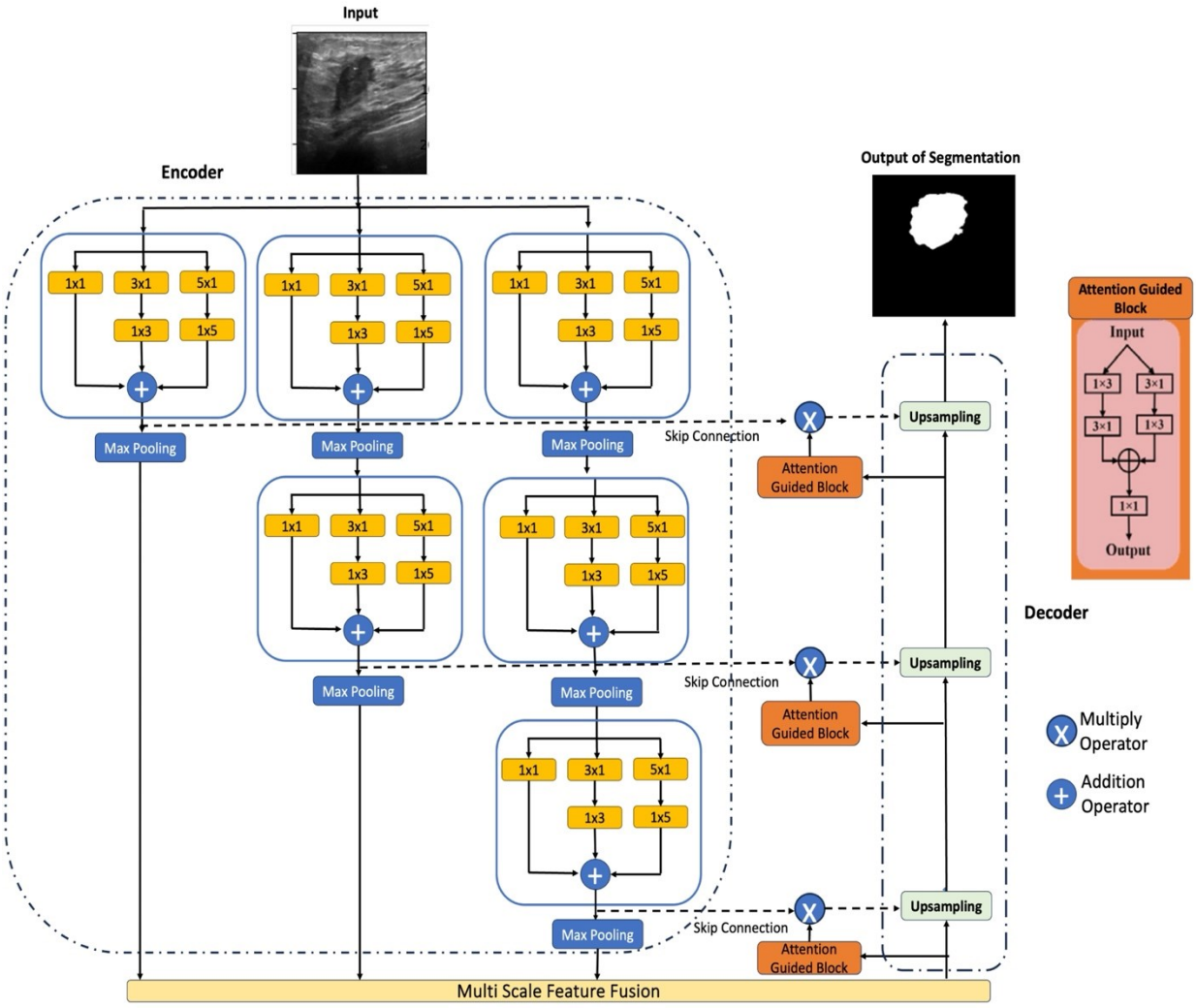


Figure 3.14: Proposed FFNet network

We propose a feature fusion-based network called FFNet, as shown in Figure 3.14, for breast ultrasound image segmentation. The proposed FFNet network consists of a multipath encoder having six context encode modules that are developed using 1×1, 3×3 and 5×5 kernels to capture information about the object of interest from breast ultrasound images. In the encoder, each $k \times k$ convolutional operation is performed using two successive one-dimensional convolutions, namely,

a $k \times 1$ kernel followed by a $1 \times k$ kernel, thereby reducing computational complexity. It should be noted that features extracted from a larger kernel are more likely to be generic and spread across the image, whereas the features extracted using a smaller kernel might extract highly local-detailed characteristics [60]. The number of filters used in the 1st, 2nd, and 3rd stages of encoder are, respectively, 34, 64, and 128. In the decoder block [61], each upsampling module integrates a spatial upsampling operation with the corresponding feature maps from the encoder via skip connections, facilitating the recovery of fine-grained spatial details and contextual information. Furthermore, the attention-guided block plays a pivotal role in adaptively emphasizing the most salient upsampled features across hierarchical levels, thereby facilitating precise boundary localization through attention-guided feature enhancement. This particular attention guided block is structured to integrate convolutions of $1 \times k$ followed by $k \times 1$ and $k \times 1$ followed by $1 \times k$, for $k = 1, 3, 5$. This size is specifically selected for its proficiency in delineating precise and localized features, thereby enhancing the visual distinction of boundaries such as those of cancer lesions or tissues. By incorporating two parallel convolutional layers, the attention guided block adaptively learns and makes adjustments for edge alignment, correcting any misalignment along the boundaries. This creates a powerful, attention-driven mechanism for the integration of features across multiple levels. Additionally, the integration of the output of the attention guided block with features from the corresponding encoder layer generates guided skip connections. These connections facilitate the fusion of feature information at both the high and low resolutions during the upsampling process. Finally, the output of the upsampling in the decoder produces the final segmentation map, isolating the objects of interest within the input image.

3.5.2 Description of Datasets

To evaluate the performance of the proposed FFNet network, we employ two publicly accessible ultrasound datasets, namely, the Breast Ultra Sound (BUS) dataset [62] and Breast UltraSound Images (BUSI) dataset [63].

The BUS dataset consists of a total of 811 ultrasound images of the breast. These images have been acquired through five varied ultrasound systems and consist of 358 benign instances and 453 malignant instances, all of which are accompanied by their respective verified ground truth annotations. This dataset has been designed by [62] for the purpose of segmenting benign and malignant tumors.

The BUSI dataset consists of ultrasound images of 600 female subjects. The data includes 133 normal, 487 benign, and 210 malignant images, with each case accompanied by accurate ground truth annotations. This dataset has been designed by [63] for the purpose of segmenting benign and malignant tumors.

3.5.3 Data Preprocessing and Training

In this preprocessing stage, we resize all the images to 256×256×3 for the two datasets that we employ. We now introduce a data augmentation method that is a contour-aware super-pixel grid mixing-based augmentation called CSGM. Our technique for enhancing data utilizes a grid-based mixing method. Our augmentation method is different from the traditional cut-mix approach [64], since we create an image by carefully blending super-pixels from a pair of images. This

augmentation technique provides the integrity of the edge details for objects within both of the images. The creation of the augmented image X_{Mix}, the corresponding mask Y_{Mix}, and the superpixel map S_{Mix} is detailed by the following equations.

$$X_{Mix} = X_1 \odot M_1 + X_2 \odot (M_2 - M_1) \tag{3.5}$$

$$Y_{Mix} = Y_1 \odot M_1 + Y_2 \odot (M_2 - M_1) \tag{3.6}$$

$$S_{Mix} = S_1 \odot M_1 + S_2 \odot (M_2 - M_1) \tag{3.7}$$

where X_1 and X_2 are two original input images, Y_1 and Y_2 are, respectively, the corresponding ground truth segmentation masks of X_1 and X_2, and S_1 and S_2 are, respectively, the two super-pixel maps of X_1 and X_2, and $\odot$ denotes the element-wise multiplication. Specifically, the augmented image X_{Mix} is produced by blending X_1 and X_2. The binary mask M_1 randomly selects superpixel regions from X_1 using the Bernoulli distribution approach, M_2 is a binary mask of all selected super-pixel regions from both X_1 and X_2, and $M_2 - M_1$ isolates the corresponding super-pixel regions from X_2, ensuring non-overlapping superpixels of the images. Y_{Mix}, the corresponding ground truth of X_{Mix}, is constructed by using Y_1 and Y_2. Finally, the super-pixel maps, S_1 and S_2 are utilized to generate a unified map S_{Mix}, maintaining structural information aligned with the mixed image regions. This augmentation strategy preserves spatial boundaries and object contours across both the images and improves the diversity of features during training of the proposed FFNet network. The process of the augmentation is shown in Figure 3.15.

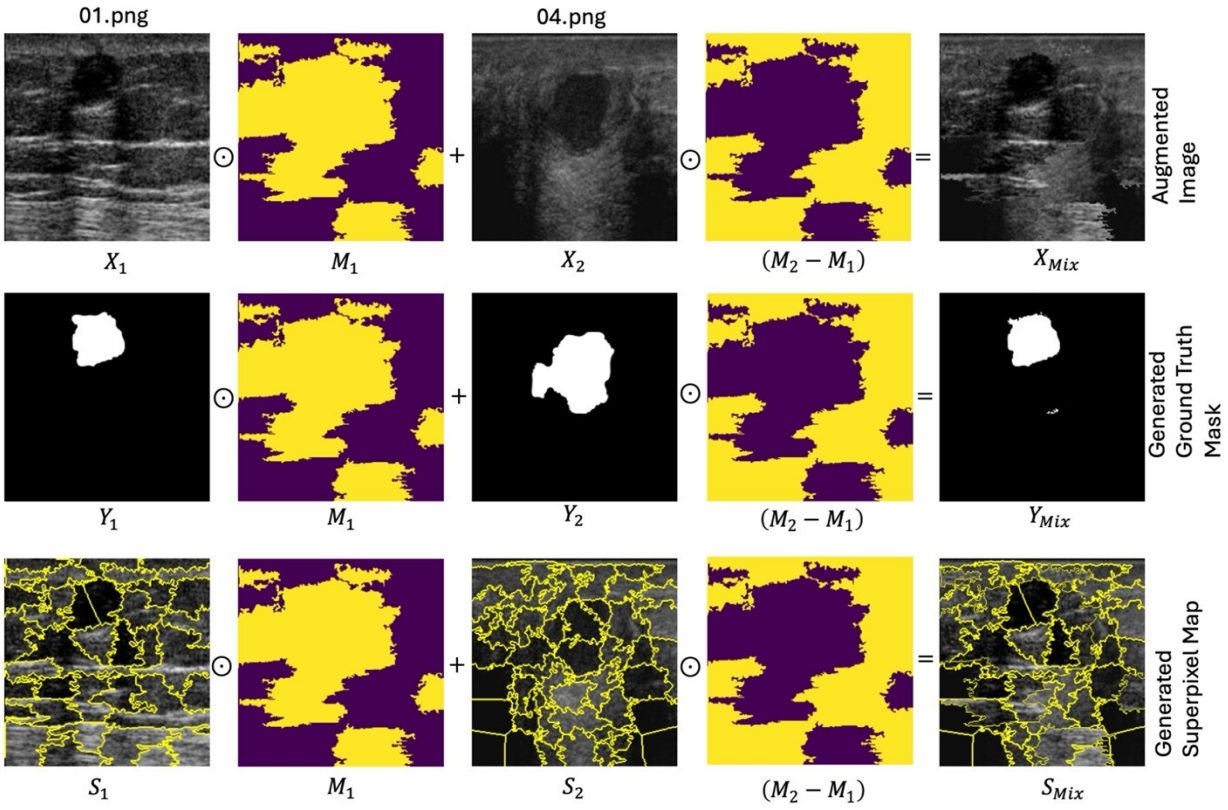


Figure 3.15: Process of contour-aware super-pixel grid mixing-based augmentation method to generate augmented image, corresponding ground truth and generated super-pixel map

For training the proposed network, a new loss function called contextual differential loss, CDL, which focuses on enhancing the segmentation performance by emphasizing the contextual differences within the image regions, is defined as

$$CDL(Y_{Pred}, Y_{true}) = L_{Seg}(Y_{Pred}, Y_{true}) + \lambda L_{DCC}(Y_{Pred}, Y_{true}) \tag{3.8}$$

where Y_{Pred} and Y_{true} denote, respectively, the predicted and ground truth segmentations, and λ is a tunable parameter used to balance between the dice loss L_{Seg} [56] and the differential contextual component L_{DCC} given by

$$L_{DCC}(Y_{Pred}, Y_{true}) = \frac{1}{N}\sum_{i=1}^{N} w_i . |D_{CM}(Y_{Pred}^i) - (Y_{true}^i)| \quad (3.9)$$

The loss function L_{DCC} employes the contextual relationships between the pixels, differentiating, in particular, between closely adjacent but different anatomical structures. It aims to provide a better understanding of the tissue boundaries and pathological regions, which are often challenging to isolate due to their similar appearance in medical images. The differential contextual mapping is employed to identify and emphasize the small differences in texture and intensity between adjacent regions. In the above equation, D_{CM} represents the contextual mapping of the images calculated using the kernel of the proposed network that captures the local context around each pixel, w_i is the weight assigned to the i-th pixel, and N is the total number of pixels in the segmentation map. The region-specific weighting, w_i, is based on the clinical relevance of the different anatomical or pathological areas.

For experimental purposes, we use Google Collaboratory with its GPU Tesla V100 that is a Jupyter notebook-based cloud service for implementing the proposed FFNet network. We also use the library packages, Matplotlib, Numpy, cv2, Sklearn, keras, and Tensorflow. We apply CSGM data augmentation before training the data. In the BUS dataset, the cancerous images are used for our experiment with 80% for training and 20% for testing. As per the BUSI dataset, we focus exclusively on the benign and malignant cases with 80% for training and 20% for testing. Adam optimizer with batch size of 8 and proposed CDL loss function are used to train the FFNet network.

3.5.4 Results of the Proposed Scheme and Comparison with that of the State-of-the-art Networks

Quantitative Results

Table 3.13: Comparisons of the performance of the proposed FFNet with that of the state-of-the-art networks using the BUS dataset

Network	DSC	IoU	Accuracy
BUS-Set [65]	0.85	0.78	0.97
RDAU-NET [66]	0.84	0.81	0.92
BUS-GAN [67]	0.87	0.77	0.90
U-Net-SA [68]	0.90	0.83	0.96
Proposed FFNet	0.94	0.92	0.98

Table 3.14: Comparisons of the performance of the proposed FFNet with that of the state-of-the-art networks using the BUSI dataset

Network	DSC	IoU	Accuracy
DDA-AttResUnet [69]	0.92	0.87	0.98
RCA-IUNet [70]	0.91	0.89	0.97
AMS-PAN [71]	0.80	0.68	0.97
Inv-UNET [72]	0.80	0.71	0.96
Proposed FFNet	0.94	0.90	0.98

Table 3.13 gives the performance, in terms of DSC, IoU and accuracy, for our FFNet network and compare the results with that of the state-of-the-art networks, BUS-Set [65], RDAU-NET [66], BUS-GAN [67], and U-Net-SA [68], in segmenting the objects of interest, namely, malignant and

benign, using the BUS dataset. Corresponding results for the BUSI dataset using the FFNet as well as the state-of-the-art networks, DDA-AttResUnet [69], RCA-IUNet [70], AMS-PAN [71], and Inv-UNET [72], are given in Table 3.14. It is seen from these two tables that the proposed FFNet network outperforms that of the state-of-the-art networks in terms all the three metrics for both datasets.

Visual Analysis

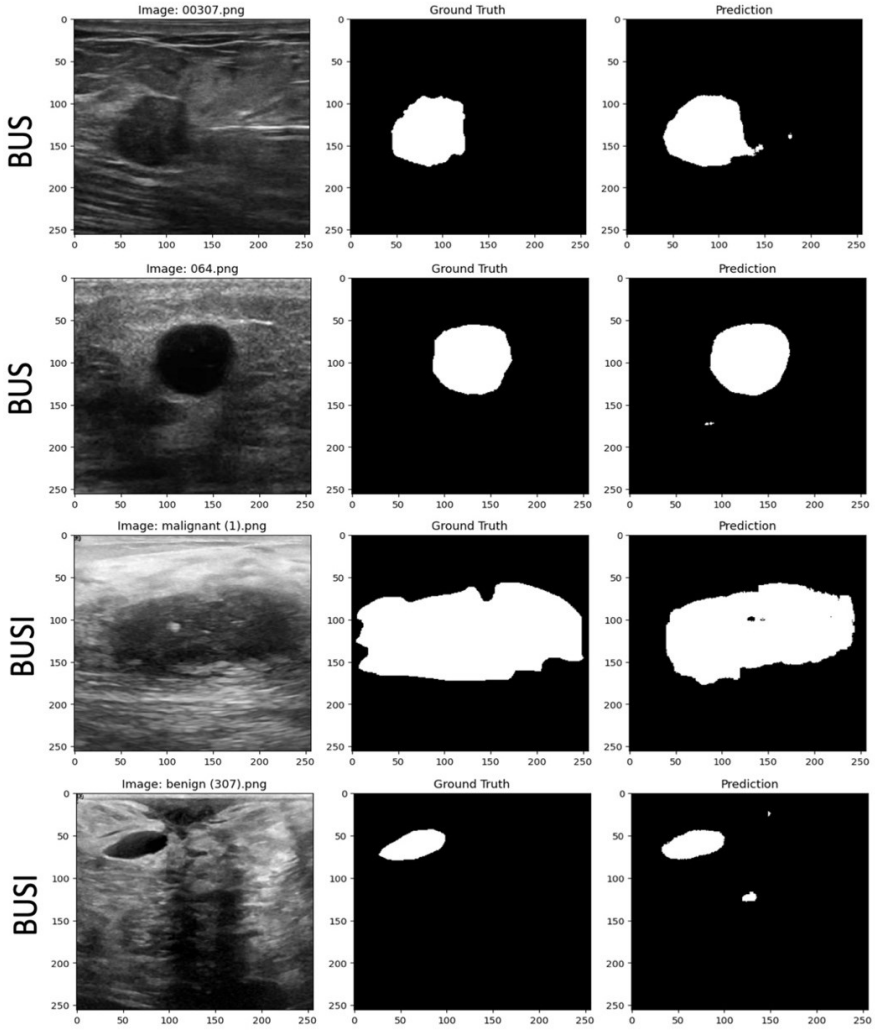


Figure 3.16: Visual illustration of the breast cancer segmentation using the proposed FFNet network with respect to ground truth

Figure 3.16 provides a visual illustration of the segmentation performance of the FFNet network on both the BUS and BUSI breast ultrasound datasets. In this figure, the first row (the second row) shows a malignant cancer image (benign cancer image) from the BUS dataset with the corresponding ground truth and the segmented image using FFNet network. The third row (the fourth row) shows a malignant cancer image (benign cancer image) from the BUSI dataset with the corresponding ground truth and the segmented image using FFNet network. It is seen from this figure that the predicted segmentation results show a strong visual alignment with the ground truth annotations, effectively capturing the tumor boundaries despite the variations in shape, intensity, and image quality.

3.6 Summary

This chapter has presented four U-Net-based architectures for medical image segmentation across various imaging modalities. The networks, MAGNet and MedSegNet, have been proposed for segmenting CT, colonoscopy, and non-mydriatic 3CCD images, with MedSegNet offering a lighter weight architecture with a performance comparable to that of the MAGNet. Further, a multi-scale attention-enhanced semi-supervised network, SSNet, has been proposed to effectively utilize both labeled and unlabeled data for segmenting brain anatomical structures of tissues in MRI images. Finally, the network FFNet has been proposed for segmenting breast cancer images, employing a convolutional neural network with a multipath encoder. It has been shown that all these four networks exhibit a performance superior to that of the existing state-of-the-art networks in terms of the standard performance metrics.

Chapter 4

A Lightweight Attention-Guided Network with Feature Recalibration for Medical Image Segmentation

4.1 Introduction

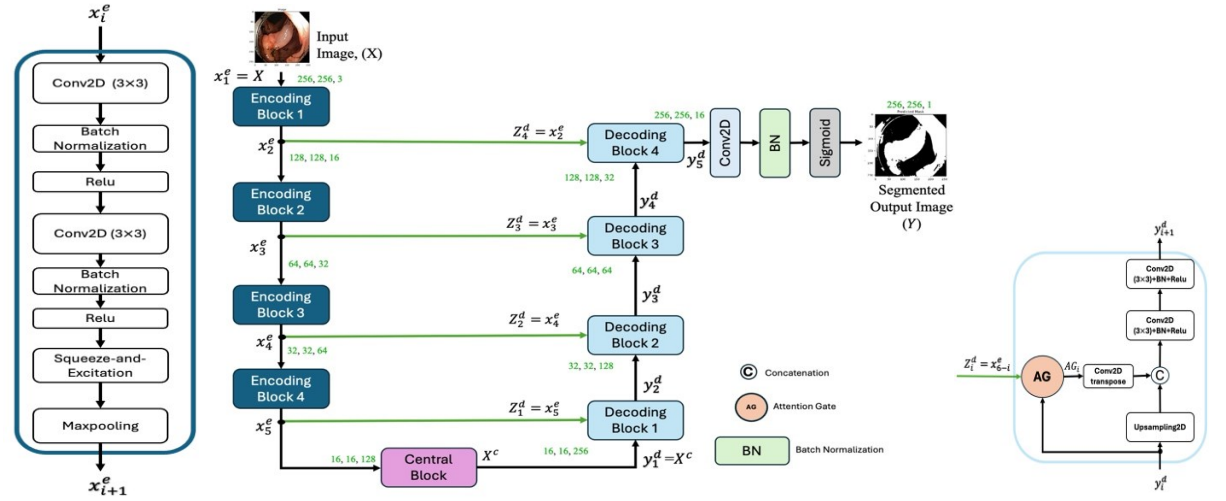
In Chapter 3, we have presented four U-Net based architectures for segmentation of medical images. While these networks have been shown to improve the performance of segmentation over that of the existing state-of-the-art networks, they require considerable amount of computational resources. To improve the performance and at the same time reduce the amount of computational resources, we propose in this chapter a lighter weight attention-guided segmentation network that is capable of providing a high segmentation accuracy across different imaging modalities, such as CT scan, colonoscopy, endoscopy and ultrasound images [73]. In addition, we also develop an optimization algorithm that provides higher segmentation performance for all used datasets. Through extensive experimentation, the proposed network is shown to outperform existing state-

of-the-art models in terms of standard performance metrics [73].

4.2 Architecture of Proposed LASegNet

In this section, we develop a lightweight attention-guided segmentation network with feature recalibration called LASegNet, using the conventional U-Net [21] shown in Figure 2.1 as the base model. Details of the LASegNet network architecture are shown in Table 4.1. The conventional U-Net contains a large number of filters and deep layers which lead to a large number of trainable parameters and increased computational cost. The proposed network is designed as a lightweight variant of the U-Net by significantly reducing the number of filters in each encoding and decoding block without compromising the performance for segmentation. Further, additional modules are introduced in the proposed network to enhance representational capability while preserving the lighter weight nature of the network. The complete architecture of the proposed network termed LASegNet is shown in Figure 4.1 (a). The network consists of four encoding and four decoding blocks. The number of filters in the encoding blocks 1, 2, 3, and 4 are, respectively, 16, 32, 64, and 128. Each encoding block contains two 3×3 convolution layers, each convolution layer being followed by batch normalization (BN) and ReLU activation. A Squeeze-and-Excitation (SE) block [74] is then added in the encoding block for channel-wise feature recalibration, allowing the network to emphasize informative features. The output of the each of the encoding blocks 1, 2, and 3 goes to a maxpooling layer to reduce spatial resolution before being fed to the next encoding block. The output of the 4^{th} encoding block goes through a maxpooling layer before it is fed into as an input of a central block that acts as a bridge between the encoder and decoder. On the decoder side, the four decoding blocks progressively upsample the feature maps. Each decoding block receives skip connections from its corresponding encoding block to recover spatial details lost during downsampling. Each skip connection in the network controls the flow of information by an

attention gate [75], allowing only the most relevant spatial information to pass to the decoding block. This attention mechanism helps in isolating the boundary accurately and reducing noise in the segmentation output. Finally, a 1×1 convolution followed by a batch normalization, and a sigmoid activation produce the final segmentation output. The following provides the details of the network architecture.



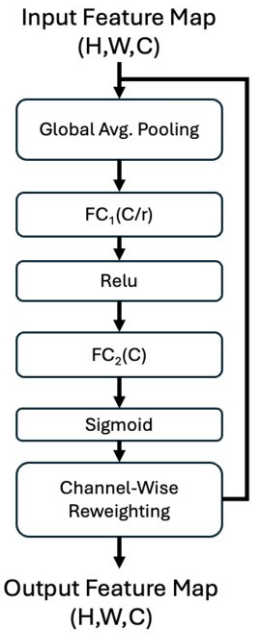
Architecture of the i^{th} Encoding Block

(i=1, 2, 3, 4)

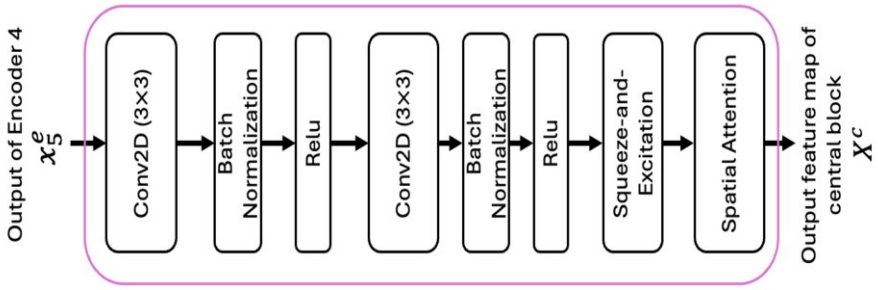
Architecture of the i^{th} Decoding Block

(i=1, 2, 3, 4)

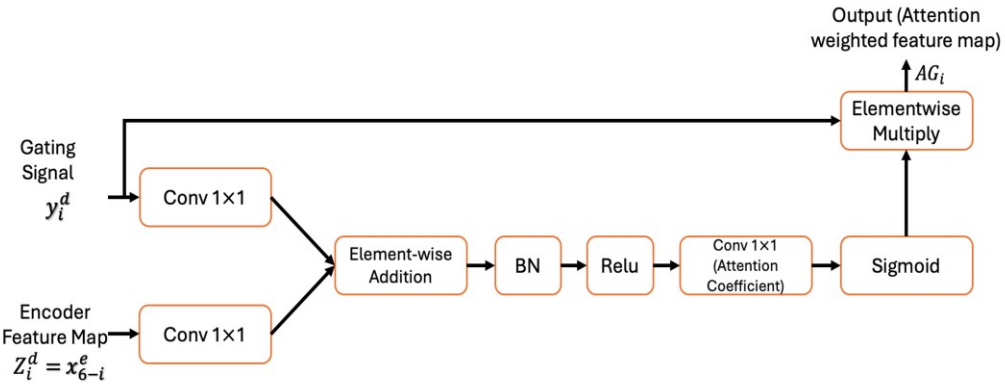
(a) Overall Architecture



(b) Squeeze-and-Excitation (SE) Block



(c) Structure of Central Block



(d) Structure of Attention Gate Block

Figure 4.1: Proposed LASegNet network

Table 4.1: Details of the LASegNet network architecture

Block	Layer	Number of input channels	Number of output channels	Input Size	Output Size
Encoding Block 1	3×3 Conv +BN+ReLu	3	16	256×256	256×256
	3×3 Conv +BN+ReLu+SE	16	16	256×256	256×256
	2×2 Maxpooling	16	16	256×256	128×128
Encoding Block 2	3×3 Conv +BN+ReLu	16	32	128×128	128×128
	3×3 Conv +BN+ReLu+SE	32	32	128×128	128×128
	2×2 Maxpooling	32	32	128×128	64×64
Encoding Block 3	3×3 Conv +BN+ReLu	32	64	64×64	64×64
	3×3 Conv +BN+ReLu+SE	64	64	64×64	64×64
	2×2 Maxpooling	64	64	64×64	32×32
Encoding Block 4	3×3 Conv +BN+ReLu	64	128	32×32	32×32
	3×3 Conv +BN+ReLu+SE	128	128	32×32	32×32
	2×2 Maxpooling	128	128	32×32	16×16
Central Block	3×3 Conv +BN+ReLu	128	256	16×16	16×16
	3×3 Conv +BN+ReLu+SE	256	256	16×16	16×16
Decoding Block 1	Upsample 2D Concat (AG + conv2D transpose)	256	128	16×16	32×32
	3×3 Conv +BN+ReLu	128	128	32×32	32×32
	3×3 Conv +BN+ReLu	128	128	32×32	32×32

Decoding Block 2	Upsample 2D Concat (AG + conv2D transpose)	128	64	32×32	64×64
	3×3 Conv +BN+ReLu	64	64	64×64	64×64
	3×3 Conv +BN+ReLu	64	64	64×64	64×64
Decoding Block 3	Upsample 2D Concat (AG + conv2D transpose)	64	32	64×64	128×128
	3×3 Conv +BN+ReLu	32	32	128×128	128×128
	3×3 Conv +BN+ReLu	32	32	128×128	128×128
Decoding Block 4	Upsample 2D Concat (AG + conv2D transpose)	32	16	128×128	256×256
	3×3 Conv +BN+ReLu	16	16	256×256	256×256
	3×3 Conv +BN+ReLu	16	16	256×256	256×256
Final Block	3×3 Conv + BN+ Sigmoid	16	1	256×256	256×256

The Squeeze-and-Excitation (SE) block in each of the encoding block as well as in the central block is designed to recalibrate channel-wise feature responses by explicitly modeling the interdependencies between channels and is shown in Figure 4.1 (b). By performing a “squeeze” operation followed by “excitation,” the SE block allows the network to focus on the most informative channels, thus enhancing the representational power of the network. This module improves the ability of the model to capture meaningful patterns and reduces the impact of less relevant features. It does so by employing global context information through a global average pooling operation, followed by a fully connected bottleneck structure that learns channel-specific

scaling factors. These scaling factors are applied to the input feature maps, producing a recalibrated output. The input feature map of size (H, W, C) is first passed through a global average pooling (GAP) layer, which compresses each feature channel to a single scalar value. This step captures the global context of each channel by summarizing the spatial information into a single value, creating a squeezed representation of the feature maps. The squeezed feature vector is then passed through two fully connected (FC) layers. First FC Layer reduces the dimensionality of the feature vector by a factor of r, where r is the reduction ratio that is set to 16. This layer captures channel dependencies in a lower-dimensional space and is followed by a ReLU activation. The second FC layer expands the reduced feature vector back to the original channel size C through another fully connected layer. This layer produces channel-wise scaling factors by applying a sigmoid activation, which normalizes the output to be between 0 and 1. The output of the second fully connected layer is used to reweight the original input feature map on a per-channel basis. In channel-wise reweighting, each channel of the input feature map is multiplied by the corresponding scaling factor, selectively increasing or decreasing each channel based on its importance. In output feature maps, the final recalibrated feature maps have the same spatial dimensions as the input (H, W, C), but the channels are now weighted according to the importance learned by the SE block.

The encoding blocks, as shown in Figure 4.1 (a), serve as the feature extraction pathway in the proposed LASegNet architecture. Each encoding block has several key components designed to progressively downsample the input image while capturing increasingly abstract and hierarchical features. The encoding block not only focuses on spatial details but also incorporates feature recalibration techniques to enhance feature representation at each level.

Let x_i^e be the input feature map to the encoding block i. The feature map from the first convolutional layer after batch normalization and ReLu activation of each encoding block, X_i^{Conv1}, is given by

$$X_i^{Conv1} = ReLu(BN(Conv_{3\times3}(x_i^e))) \tag{4.1}$$

The feature map from the second convolutional layer after batch normalization and ReLu activation of each encoding block, X_i^{Conv2}, is given by

$$X_i^{Conv2} = ReLu(BN\left(Conv_{3\times3}(X_i^{Conv1})\right)) \tag{4.2}$$

These convolutions allow the encoding block to learn spatial patterns, such as edges, textures, and shapes, from the input. The number of filters is progressively increased as the encoding block goes deeper in the network, enabling it to capture local to global features. The output of X_i^{Conv2} is fed into Squeeze-and-Excitation block, described above. The output, X_i^{se}, of the Squeeze-and-Excitation block is given by

$$X_i^{se} = SE(X_i^{Conv2}) \tag{4.3}$$

The SE block is used to recalibrate the feature maps in a channel-wise manner. This process ensures that the network focuses on the most relevant input features, which is particularly valuable in complex tasks like medical image segmentation. Finally, the output of the SE block is passed through a 2×2 MaxPooling layer to obtain

$$x_{i+1}^e = MaxPooling_{2\times2}(X_i^{se}) \tag{4.4}$$

where x_{i+1}^{e} is the output feature map from the encoding block after downsampling. MaxPooling is typically applied to reduce the spatial dimensions of the feature maps, enabling the network to capture larger receptive fields at deeper levels.

The architecture of the central block of Figure 4.1 (a) is shown in Figure 4.1 (c); it combines multi-level convolutional operations with advanced attention mechanisms to refine feature maps, making them more informative. The inclusion of the SE Block ensures that the network can prioritize the most relevant channels, while the spatial attention block further enhances the spatial importance of the features. This dual attention mechanism improves the capability of the network to capture both the local and global context, making it particularly well-suited for segmentation task. The architecture of the central block is similar to that of the encoding block; however, it incorporates a spatial attention block for enhancing the features and does not contain a MaxPooling layer.

The architecture of the attention gate (AG) in the decoder side in Figure 4.1 (a) is shown in Figure 4.1 (d); it is designed to control the flow of information by paying more attention to the important features which come from the encoding block. By employing a gating signal from the deeper layers of the decoder, the attention gate allows the network to capture the most relevant information of the encoder feature map, which contains spatial and low-level features extracted during the downsampling process. Let the input to the i^{th} decoding block be denoted as y_i^d. The input y_1^d is the output X^c of the central block, while for the other decoding blocks the input is the output of the previous decoding block. The first input y_1^d serves as the gating signal for the corresponding attention gate. The second input to the attention gate, denoted by Z_i^d, is obtained from the output feature map of the x_{6-i}^{e} encoding block. Both the gating signal y_i^d and the input feature map of the attention gate Z_i^d pass through 1×1 convolution layers, thus reducing the spatial dimensions of channel. This process ensures that the dimensionality of the features is aligned, allowing them to

be compared effectively during the attention map calculation. Let $\Phi(y_i^d)$ and $\Theta(Z_i^d)$ represent the outputs of the 1×1 convolution on y_i^d and Z_i^d, respectively. The outputs $\Phi(y_i^d)$ and $\Theta(Z_i^d)$ are combined using element-wise addition to obtain F, which merges the encoder feature map and the gating signal, thus allowing the network to compute both the local and global context. Batch normalization is applied on the output of the element-wise addition to stabilize the training process and ensure that the network converges efficiently. This is followed by a ReLU activation operation to introduce non-linearity, ensuring that the network can learn complex non-linear relationships between the features. Thus, the output after the batch normalization and ReLu activation is given by

$$F_{relu} = ReLu(BN(F)) \tag{4.5}$$

A 1×1 convolution is performed on F_{relu}, followed by a sigmoid activation operation to obtain an attention coefficient with a value between 0 and 1. This output, denoted by $\psi(F_{relu})$, is elementwise multiplied with the gating signal y_i^d to produce an attention-weighted feature map y_{i+1}^{attn}. This attention-weighted feature map which contains the most relevant spatial information is then fed to the decoding block for further processing, enabling the decoder to produce more accurate segmentation results.

The decoder, as shown in Figure 4.1 (a), is an important component in the network's upsampling pathway. The primary function of the decoder is to restore the spatial resolution of the feature maps progressively, while integrating high-level abstract features from the central block and fine-grained details from the encoder via skip connections. The decoder uses attention gates to focus on the most relevant regions of the input feature maps, enhancing the performance of the proposed LASegNet network. Each decoding block upsamples the feature map to a higher resolution for

restoring the original input image dimensions for accurate localization of segmentation boundaries. The input feature map y_i^d is first upsampled in each of the decoding blocks to obtain the corresponding output, y_i^{up}. This upsampled decoder feature map, y_i^{up}, is concatenated with the output of the attention gate y_{i+1}^{attn} to obtain y_i^{concat}. This operation merges the high-level semantic features from the decoding block with the low-level, spatially detailed features from the encoding block, enabling the network to extract both local and global features. After the concatenation, in each decoding block, two 3×3 convolution operations are performed on y_i^{concat}, each convolution layer being followed by batch normalization and ReLU activation. The output of each decoding block, y_{i+1}^d, provides a reconstructed feature map with a spatial resolution which is twice that of the previous decoding block. The output from the fourth decoding block goes through a 1×1 convolution layer followed by batch normalization and a sigmoid activation function to produce the final segmentation output.

4.3 Description of Datasets

To evaluate the performance of the proposed LASegNet network, we employ six publicly accessible datasets, namely, CVC-ClinicDB [76], CVC-ColonDB [77], Kvasir-SEG [78], ETIS-LaribPolypDB [79], MosMed COVID-19 CT Scans [80] and DDTI Ultrasound [81].

CVC-ClinicDB [76]: The CVC-ClinicDB dataset consists of colonoscopy images for polyp detection and segmentation. It contains a total of 612 images of real colonoscopies, along with corresponding manually annotated ground truth masks. The dataset presents challenges due to variations in location conditions, polyp shapes, and sizes. It is one of the most-commonly used datasets for evaluating polyp segmentation in gastrointestinal colonoscopy.

CVC-ColonDB [77]: The CVC-ColonDB dataset contains 300 colonoscopy images, each paired with a binary ground truth mask isolating the presence of polyps. This dataset is of a higher complexity than CVC-ClinicDB, as it includes harder-to-detect polyps with unclear boundaries and varying sizes. CVC-ColonDB is commonly used to test the robustness of segmentation models on more challenging colonoscopic cases, making it ideal for evaluating generalization performance. This dataset has been designed by [77] for the purpose of segmenting polyps from colonoscopy images.

Kvasir-SEG Dataset [78]: The Kvasir-SEG dataset focuses on pixel-level segmentation of polyps. It includes 1,000 endoscopy images with corresponding ground truth masks. The dataset features high variability in polyp appearances, including different polyp shapes, colors, and textures. This variability makes Kvasir-SEG a particularly challenging benchmark for evaluating the adaptability of segmentation models to diverse clinical cases. This dataset has been designed by [78] for the purpose of segmenting polyps from endoscopy images.

ETIS-LaribPolypDB [79]: The ETIS-LaribPolypDB dataset contains 196 endoscopy images of polyps with its corresponding ground truth, characterized by highly irregular boundaries and challenging lighting conditions. This dataset presents challenges for accurate polyp segmentation due to low contrast and complex textures in the images. This dataset has been designed by [79] for the purpose of segmenting polyps from endoscopy images.

MosMed COVID-19 CT Scans [80]: The MosMed dataset contains 2729 chest CT scans of patients with confirmed COVID-19 with the corresponding ground truth masks, indicating regions affected by infection. The segmentation task involves detecting lesions caused by COVID-19 in the lungs. This dataset provides a critical benchmark for lung lesion segmentation in CT images. This dataset has been designed by [80] for the purpose of segmenting lung infection from CT scan images.

DDTI Ultrasound Dataset [81]: The Digital Database for Thyroid Imaging (DDTI) dataset consists of 637 ultrasound images of patients with thyroid ultrasound. Each image is paired with an expert-annotated mask that identifies the thyroid nodule or abnormalities. Ultrasound images pose unique challenges due to noise, low contrast, and anatomical variability. This dataset is often used to evaluate segmentation performance in ultrasound imaging, where segmentation of thrombus regions is crucial for the accurate diagnosis of thyroid conditions. This dataset has been designed by [81] for the purpose of segmenting thyroid nodules from ultrasound images.

The above information is summarized in Table 4.2.

Table 4.2: Summary of datasets used in testing the proposed LASegNet network for segmentation

Dataset	**Modality**	**Images**	**Resolution**	**Proportion used for training, validation and testing**
CVC-ClinicDB [76]	Colonoscopy	612	384×288	70%, 15%, 15%
CVC-ColonDB [77]	Colonoscopy	300	574×500	70%, 15%, 15%
Kvasir-SEG [78]	Endoscopy	1000	Varies	70%, 15%, 15%
ETIS-LaribPolypDB [79]	Endoscopy	196	225×966	70%, 15%, 15%
Mosmed COVID-19 CT Scans [80]	Computed Tomography (CT)	2729	512×512	70%, 15%, 15%
DDTI [81]	Ultrasound Imaging	637	1024×768	70%, 15%, 15%

4.4 Data Preprocessing

All the images in the above-mentioned datasets are first resized to 256×256 pixels. The image resizing and preprocessing are handled based on the file format associated with each dataset. The dataset is first loaded by reading the images and corresponding ground truth masks from their respective directories. The images are normalized to the range [0, 1] by dividing the pixel values

by 255. In order to make the training data more diverse, several augmentation techniques are applied for the different modalities. These include *horizontal* and *vertical flips* [82], which help the network to learn from different orientations of the images, which is crucial for medical images where the orientations can vary, *ColorJitter* augmentation [83] that introduces variations in brightness, contrast, saturation, and hue, simulating different lighting conditions and color changes. In addition, *Affine* transformation [84] is also utilized which to mimic variations in size, position, and shape, thus helping the model to become more invariant to these spatial transformations.

4.5 Training and Validation

As mentioned in Table 4.2, in each dataset 70% is utilized for training, 15% for validation and the remaining 15% for testing. For training the data, we use the Contextual Differential Loss introduced in Chapter 3 as the loss function. To optimize the network, we propose an optimizer called fAdamR, which is described below.

The fAdamR optimizer is an optimization algorithm that combines the strengths of both the *Adam* [26] and RMSProp [29], two widely used optimization algorithm used in medical image segmentation. *Adam* optimizer employs adaptive learning rates based on the estimates of the first and second moments, incorporates momentum and applies bias correction for improved stability and convergence. In contrast, RMSProp focuses on adaptive learning rates derived solely from the estimates of the second moment, excelling in tasks with non-stationary objectives but lacking momentum and bias correction. By fusing the strengths of these two optimizers, the proposed fAdamR benefits from the *Adam*'s bias correction for momentum and adaptive learning rates, while also benefiting from RMSProp's ability to reduce oscillations in non-convex optimization. The

proposed fAdamR employs an adaptive parameter update rule that balances the contributions from the two optimizers, providing good performance in segmentation. The algorithm of the fAdamR optimizer is given below.

Let the learning rate be denoted by η, the exponential decay rates for moment estimates by $\beta_1$and β_2, and a small constant to prevent division by zero by ϵ.

Algorithm 4.1: fAdamR optimizer

Initialize:

Initialize first moment vector $\boldsymbol{m_0 = 0}$

Initialize second moment vector $\boldsymbol{v_0 = 0}$

Initialize time step $\boldsymbol{t_0 = 0}$

For each iteration $\boldsymbol{t}$:

1. Increment time step:
$$\boldsymbol{t \leftarrow t + 1}$$
2. Compute gradients:
$$\boldsymbol{g_t \leftarrow \nabla_\theta f_t(\theta_{t-1})}$$
3. Update biased 1st moment estimate:
$$\boldsymbol{m_t \leftarrow \beta_1 . m_{t-1} + (1 - \beta_1) . g_t}$$
4. Update biased 2nd moment estimate:
$$\boldsymbol{v_t \leftarrow \beta_2 . v_{t-1} + (1 - \beta_2) . g_t^2}$$
5. Calculate the bias-corrected 1st moment estimate:
$$\boldsymbol{\widehat{m_t} \leftarrow \frac{m_t}{1 - \beta_1^t}}$$
6. Calculate the bias-corrected 2nd moment estimate:
$$\boldsymbol{\widehat{v_t} \leftarrow \frac{v_t}{1 - \beta_2^t}}$$
7. Compute *Adam*:
$$\boldsymbol{Adam \leftarrow \eta . \frac{\widehat{m_t}}{\sqrt{(\widehat{v_t} + \varepsilon)}}}$$
8. Compute RMSProp:
$$\boldsymbol{RMSProp \leftarrow \eta . \frac{g_t}{\sqrt{v_t + \varepsilon}}}$$
9. Combine updates:
$$\boldsymbol{Fusion\_update \leftarrow r . RMSProp + (1 - r) Adam}$$
10. Update parameters:
$$\boldsymbol{\theta_t \leftarrow \theta_{t-1} - Fusion\_update}$$

End For

In the above Algorithm, the time step t is incremented by 1, and the gradient g_t of the loss function f_t with respect to the parameters θ is computed in Step 2, namely, computing the gradients. In Steps 3 and 4, the 1st and 2nd moment estimates, m_t and v_t, are updated using exponential decay factors β_1 and β_2. In Steps 5 and 6, the bias-corrected 1st moment estimate ($\widehat{m_t}$) and the bias-corrected 2nd moment estimate ($\widehat{v_t}$) are computed to adjust the bias introduced during initialization. In Step 7, the adaptive moment estimation, *Adam*, is computed using the bias-corrected 1st moment estimate, while in Step 8, the RMSProp update is computed using the bias-corrected 2nd moment estimate. In Step 9, the updates from *Adam* and RMSProp are combined using the fusion ratio r, which we have chosen for our experiments to be 0.5. In Step 10, the parameters θ are updated using the combined $Fusion\_update$ in the last step. This algorithm balances the adaptive learning rate feature of RMSProp with the momentum and bias correction of Adam, providing an efficient optimization process in the segmentation task.

Training and Validation Curve

The proposed network is implemented using Keras with TensorFlow in a Google Colab environment that supports an NVIDIA Tesla K80 GPU. All training, validation, and testing are conducted in this consistent environment. The network architecture is developed and debugged using TensorFlow [85] and Keras [86], with the proposed optimization algorithm, fAdamR, employed for training. The training is conducted over 75 epochs with an initial learning rate of 1e-4, and the batch size set to 16. Figure 4.2 depicts the training and validation loss curves for our proposed network for all the six datasets, evaluating its performance in terms of epoch vs. loss. Minimal overfitting is observed, as indicated by the close alignment between the training and

validation loss curves for all the six datasets, highlighting the robustness of the proposed network for segmentation of medical images of various modalities.

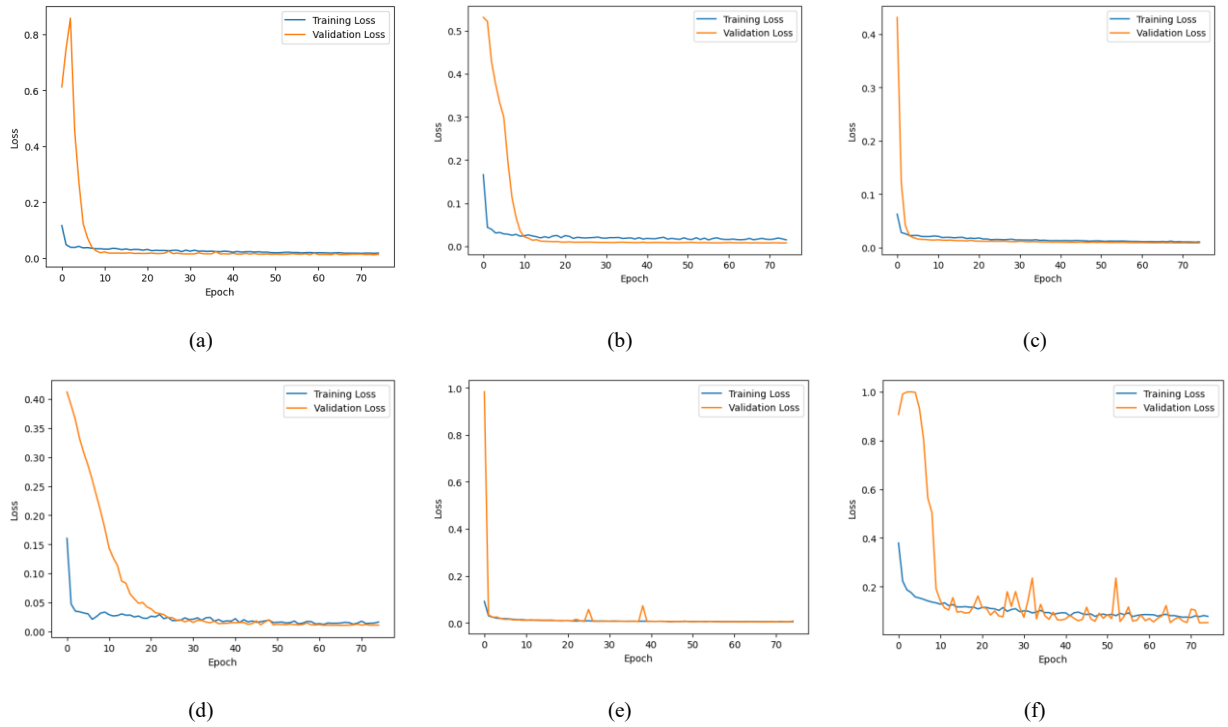


Figure 4.2: Evaluation of training and validation performance of our proposed network for (a) CVC-ClinicDB, (b) CVC-ColonDB, (c) Kvasir-SEG, (d) ETIS-LaribPolypDB, (e) Mosmed COVID-19 CT scans, and (f) DDTI in terms of epoch vs. loss

4.6 Results and Comparisons

In this section, we evaluate the performance of the proposed scheme and compare it with that of the state-of-the-art schemes, namely, the schemes employing U-Net [21], Attention U-Net [87], Dense U-Net [88], HyperDense-Net [89], TransUNet [90], and MISSFormer [91], for all the six

datasets, namely, CVC-ClinicDB, CVC-ColonDB, Kvasir-SEG, ETIS-LaribPolypDB, Mosmed COVID-19 CT scans.

Evaluation of the Effectiveness of the Proposed Scheme through an Ablation Study

Table 4.3: Results of proposed network and ablation study on all experimental datasets

Dataset	Metrics	Proposed Network	Proposed Network without Attention Gates	Proposed Network without SE blocks	Proposed Network without SE blocks and Attention Gates
CVC-ClinicDB	Dice Coeff.	0.93	0.84	0.86	0.82
	IoU	0.86	0.74	0.77	0.72
	Precision	0.95	0.85	0.87	0.80
	Recall	0.92	0.82	0.84	0.79
	Avg.	0.91	0.81	0.83	0.78
CVC-ColonDB	Dice Coeff.	0.90	0.82	0.83	0.81
	IoU	0.89	0.70	0.72	0.65
	Precision	0.92	0.83	0.84	0.83
	Recall	0.90	0.80	0.81	0.79
	Avg.	0.90	0.78	0.80	0.77
Kvasir-SEG	Dice Coeff.	0.94	0.83	0.85	0.80
	IoU	0.87	0.72	0.74	0.72
	Precision	0.94	0.84	0.86	0.80
	Recall	0.92	0.81	0.83	0.75
	Avg.	0.91	0.80	0.82	0.76
ETIS-LaribPolypDB	Dice Coeff.	0.91	0.80	0.82	0.76
	IoU	0.88	0.68	0.70	0.66
	Precision	0.90	0.81	0.83	0.82
	Recall	0.88	0.79	0.80	0.77
	Avg.	0.89	0.77	*0.79*	*0.75*
Mosmed COVID-19 CT scans	Dice Coeff.	0.88	0.81	0.83	0.80
	IoU	0.86	0.69	0.71	0.67
	Precision	0.90	0.83	0.84	0.80
	Recall	0.89	0.79	0.80	0.76
	Avg.	*0.88*	*0.78*	0.80	*0.75*
DDTI	Dice Coeff.	0.87	0.82	0.83	0.81
	IoU	0.80	0.70	0.71	0.72
	Precision	0.89	0.83	0.84	0.81
	Recall	0.90	0.80	0.81	0.76
	Avg.	0.86	0.79	0.80	0.77
Overall Avg.		0.89	0.79	0.81	0.76

Table 4.3 gives the results of the ablation study on the effectiveness of the attention gates and squeeze-and-excitation (SE) blocks on the performance of our proposed LASegNet network in terms of dice coefficient, IoU, precision and recall. The performance is evaluated on the six

different medical imaging datasets by obtaining the values of the four-segmentation metrics for each dataset. The third column of the table provides the performance of the proposed network. The fourth, fifth and sixth columns provide the performance when the attention gates, SE blocks, and both the attention gates and SE blocks are, respectively, removed from the proposed network. This table also provides for each of the datasets the performance averaged over all the four metrics, as well as the overall averaged performance that is the performance averaged over all the metrics and all the datasets. It is observed from the table that regardless of keeping the two modules in the proposed network or removing either or both the modules from the network, the best average performance is obtained on the CVC-ClinicDB dataset, and the lowest performance is obtained on the Mosmed COVID-19 CT scans dataset. It is also seen from the table that for each dataset, the averaged performance over all the metrics deteriorates significantly when either of the two modules is removed from the proposed LASegNet network, and more so, when both modules are removed. This observation also holds in terms of the overall average performance.

Quantitative Results

Table 4.4 presents a comparison of training times of the proposed network and the state-of-the-art networks, namely, U-Net [21], Attention U-Net [87], Dense U-Net [88], HyperDense-Net [89], TransUNet [90], and MISSFormer [91] using two optimizers, Adam and fAdamR, on the six different medical imaging datasets. It is seen from this table that regardless of the network and regardless of the datasets used, the training times using fAdamR is slightly higher than the training times using Adam optimizer and this additional training time using fAdamR results from the operation of RMSProp optimizer that is integrated to this optimizer. It is observed from the table that the proposed LASegNet network exhibits significantly lower training times compare to the

state-of-the-art networks for all the datasets. For example, for the CVC-ClinicDB dataset, training time of the proposed network is 4.0-hours using fAdamR optimizer, which is significantly lower training time taken by the other networks, which ranges between 4.9 and 13 hours.

Table 4.4: Comparison of proposed network with state-of-the-art networks in terms of training time

Dataset	**Optimizer**	**LASegNet (Proposed)**	**U-Net [21]**	**Attention U-Net [87]**	**Dense U-Net [88]**	**HyperDense-Net [89]**	**TransUNet [90]**	**MISSFormer [91]**
CVC-ClinicDB	Adam	~3.96 hrs	~7.5 hrs	~5.5 hrs	~4.7 hrs	~9.3 hrs	~12.7 hrs	~8.2 hrs
	fAdamR	~4 hrs	~7.7 hrs	~5.6 hrs	~4.9 hrs	~9.4 hrs	~13 hrs	~8.5 hrs
CVC-ColonDB	Adam	~3.73 hrs	~7.45 hrs	~5.2 hrs	~4.55 hrs	~9.1 hrs	~12.5 hrs	~8.1 hrs
	fAdamR	~3.81 hrs	~7.48 hrs	~5.28 hrs	~4.67 hrs	~9.22 hrs	~12.62 hrs	~8.15 hrs
Kvasir-SEG	Adam	~3.99 hrs	~7.8 hrs	~5.7 hrs	~4.92 hrs	~9.6 hrs	~13.1 hrs	~8.6 hrs
	fAdamR	~4.1 hrs	~7.92 hrs	~5.77 hrs	~5.21 hrs	~9.8 hrs	~13.7 hrs	~8.93 hrs
ETIS-LaribPolypDB	Adam	~22.95 mins	~1.4 hrs	~47 mins	~33 mins	~1.9 hrs	~3.2 hrs	~1.6 hrs
	fAdamR	~23.1 mins	~1.72 hrs	~52 mins	~37 mins	~2.22 hrs	~3.46 hrs	~1.75 hrs
Mosmed COVID-19 CT Scans	Adam	~3.98 hrs	~7.72 hrs	~5.8 hrs	~4.9 hrs	~9.5 hrs	~13.1 hrs	~8.6 hrs
	fAdamR	~4.3 hrs	~7.88 hrs	~5.92 hrs	~5.22 hrs	~9.75 hrs	~13.4 hrs	~8.9 hrs
DDTI	Adam	~3.96 hrs	~7.4 hrs	~5.4 hrs	~4.7 hrs	~9.2 hrs	~12.5 hrs	~8.1 hrs
	fAdamR	~4.21 hrs	~7.6 hrs	~5.6 hrs	~5.1 hrs	~9.4 hrs	~12.9 hrs	~8.5 hrs
Avg. time for Adam		**~3.34 hrs**	~6.33 hrs	~4.73 hrs	***~4.05 hrs***	~8.10 hrs	~11.18 hrs	~7.2 hrs
Avg. time for fAdamR		**~3.47 hrs**	~6.72 hrs	~4.84 hrs	***~4.29 hrs***	~8.30 hrs	~11.51 hrs	~7.46 hrs

Table 4.5 provides the performance results, in terms of the dice coefficient, IoU, precision and recall, of all the seven segmentation schemes including the proposed LASegNet network when applied using the two optimizers, Adam and fAdamR, to the six datasets. It also gives the number of parameters used in the networks and the testing time for each of the schemes. It is observed from this table that regardless of the metrics, optimizers, and datasets, the performance of the proposed scheme is the best and that of MISSFormer and TransUNet are, generally, the second and third best, respectively. The superiority in performance of the proposed scheme over that of the schemes providing the second and third best performance is achieved by using only a fraction of the number parameters employed by the latter two schemes. Note that the proposed network requires 2.2 M, whereas MISSFormer and TransUNet require 42.46 M and 105 M, respectively. Even the network of the scheme DenseUNet that uses 4 M parameters, which is almost two times the number of the

parameters used by the network of the proposed scheme has the testing time 88 ms compared to 48 ms that of the proposed scheme. A comparison of the values of the performance metrics provided in Table 4.5 shows the use of the fAdamR optimizer results in a slightly improved performance over that using fAdamR optimizer. The fact that all the schemes benefit from using fAdamR indicates that this optimizer has superiority over Adam, but at the expense of a slightly larger training time.

Table 4.5 also provides the averaged performance over all the four metrics and standard deviation for each of the datasets individually as well as over all the datasets. The fact that the value of the overall standard deviation 0.023 for the performance of the proposed scheme is the lowest among that of the performance of all the schemes indicates that the proposed scheme for segmentation is robust across different modalities medical images.

Table 4.5: Comparison of the proposed network with the state-of-the-art networks for all the six datasets

Dataset	**Optimizer**	**Metrics**	**LASegNet (Proposed)**	**U-Net [21]**	**Attention U-Net [87]**	**Dense U-Net [88]**	**HyperDense-Net [89]**	**TransUNet [90]**	**MISSFormer [91]**
		No. of parameters	**2.2 M**	31.04 M	6.4 M	***4 M***	60 M	105 M	42.46 M
CVC-ClinicDB	Adam	Dice Coeff.	**0.91**	0.80	0.82	0.82	0.84	0.85	***0.87***
		IoU	**0.83**	0.72	0.74	0.70	0.75	0.75	***0.78***
		Precision	**0.90**	0.80	0.85	0.85	0.85	0.87	***0.89***
		Recall	**0.88**	0.80	0.81	0.83	0.82	0.83	***0.85***
		Avg. (Std.)	**0.88 (0.031)**	0.78 (0.035)	0.81 (0.040)	0.80 (0.059)	0.82 ***(0.039)***	0.83 (0.046)	***0.85*** (0.042)
	fAdamR	Dice Coeff.	**0.93**	0.83	0.85	0.85	0.86	0.87	***0.88***
		IoU	**0.86**	0.74	0.76	0.73	0.77	0.79	***0.80***
		Precision	**0.95**	0.86	0.88	0.89	0.87	0.89	***0.90***
		Recall	**0.92**	0.83	0.84	0.85	0.85	0.86	***0.87***
		Avg. (Std.)	**0.91 (0.032)**	0.82 (0.045)	0.83 (0.044)	0.83 (0.060)	0.84 (0.039)	0.85 ***(0.038)***	***0.86 (0.038)***
	-	Test Time	**~48 ms**	~678ms	~140ms	***~88ms***	~1.1s	~2.3s	~927ms
CVC-ColonDB	Adam	Dice Coeff.	**0.87**	0.78	0.80	0.81	0.82	0.82	***0.84***
		IoU	**0.85**	0.63	0.69	0.69	0.70	0.72	***0.80***
		Precision	**0.90**	0.80	***0.83***	0.81	***0.83***	***0.83***	***0.80***
		Recall	**0.88**	0.77	0.78	0.80	0.81	0.81	***0.83***
		Avg. (Std.)	**0.88 (0.018)**	0.75 (0.067)	0.78 (0.052)	0.78 (0.051)	0.79 (0.052)	0.80 ***(0.044)***	***0.81*** **(0.018)**
	fAdamR	Dice Coeff.	**0.90**	0.81	0.83	0.82	0.84	***0.85***	***0.85***
		IoU	**0.89**	0.68	0.70	0.69	0.72	0.74	***0.82***

Dataset	Optimizer	Metrics	LASegNet (Proposed)	U-Net [21]	Attention U-Net [87]	Dense U-Net [88]	HyperDense-Net [89]	TransUNet [90]	MISSFormer [91]
		No. of parameters	**2.2 M**	31.04 M	6.4 M	***4 M***	60 M	105 M	42.46 M
		Precision	**0.92**	0.82	0.84	0.82	0.85	***0.86***	***0.86***
		Recall	**0.90**	0.79	0.81	0.84	0.83	0.84	***0.85***
		Avg. (Std.)	**0.90 (0.011)**	0.78 (0.056)	0.80 (0.056)	0.79 (0.060)	0.81 (0.052)	0.82 (0.048)	***0.85 (0.015)***
		Test time	**~43ms**	~606ms	~103.1ms	***~76ms***	~1.3s	~2.5s	~1.1s
Kvasir-SEG	Adam	Dice Coeff.	**0.92**	0.79	0.81	0.83	0.85	0.83	***0.86***
		IoU	**0.85**	0.70	0.70	0.69	0.71	0.73	***0.75***
		Precision	**0.92**	0.79	0.82	0.82	***0.85***	***0.85***	0.83
		Recall	**0.90**	0.76	0.79	0.79	0.81	0.83	***0.84***
		Avg. (Std.)	**0.90 (0.029)**	0.76 (0.037)	0.78 (0.047)	0.78 (0.055)	0.81 (0.057)	0.81 (0.047)	***0.82 (0.042)***
	fAdamR	Dice Coeff.	**0.94**	0.82	0.84	0.84	0.85	0.86	***0.87***
		IoU	**0.87**	0.71	0.73	0.70	0.74	0.76	***0.79***
		Precision	**0.94**	0.84	0.85	0.83	0.86	0.87	***0.88***
		Recall	**0.92**	0.80	0.82	0.80	0.83	***0.85***	***0.85***
		Avg. (Std.)	**0.91 (0.029)**	0.79 (0.049)	0.81 (0.047)	0.79 (0.055)	0.82 (0.047)	0.83 (0.044)	***0.85 (0.035)***
		Test time	**~47ms**	~663ms	~136ms	***~85ms***	~1.1s	~2.2s	~906ms
ETIS-LaribPolypDB	Adam	Dice Coeff.	**0.88**	0.75	0.76	0.75	0.79	0.80	***0.81***
		IoU	**0.85**	0.60	0.63	0.66	0.65	0.67	***0.76***
		Precision	**0.88**	0.75	0.75	***0.82***	0.80	***0.82***	***0.82***
		Recall	**0.85**	0.73	0.76	0.79	0.76	0.79	**0.85**
		Avg. (Std.)	**0.87 (0.015)**	0.71 (0.063)	0.73 (0.055)	0.76 (0.060)	0.75 (0.059)	0.77 (0.059)	0***.81 (0.032)***
	fAdamR	Dice Coeff.	**0.91**	0.78	0.79	0.77	0.80	0.81	***0.85***
		IoU	**0.88**	0.63	0.65	0.67	0.67	0.69	***0.85***
		Precision	**0.90**	0.79	0.80	0.83	0.82	0.83	***0.84***
		Recall	**0.88**	0.76	0.77	0.79	0.78	0.80	***0.81***
		Avg. (Std.)	**0.89 (0.013)**	0.74 (0.064)	0.75 (0.060)	0.77 (0.059)	0.76 (0.058)	0.78 (0.055)	***0.84 (0.016)***
		Test time	**~44ms**	~620ms	~126ms	***~79ms***	~1.2s	~2.1s	845ms
Mosmed COVID-19 CT Scans	Adam	Dice Coeff.	**0.83**	0.78	0.78	0.80	0.81	***0.82***	***0.82***
		IoU	**0.75**	0.64	0.66	0.66	0.68	0.70	***0.73***
		Precision	**0.84**	0.79	0.80	0.80	0.82	0.82	***0.83***
		Recall	**0.82**	0.79	0.78	0.79	0.79	***0.80***	***0.80***
		Avg. (Std.)	**0.81 (0.035)**	0.75 (0.064)	0.76 (0.056)	0.76 (0.059)	0.77 (0.056)	0.79 (0.049)	0.79 (0.039)
	fAdamR	Dice Coeff.	**0.88**	0.80	0.81	0.82	0.82	***0.83***	***0.83***
		IoU	**0.86**	0.66	0.68	0.67	0.70	0.72	***0.73***
		Precision	**0.90**	0.81	0.83	0.84	0.83	0.84	***0.85***
		Recall	**0.89**	0.79	0.80	***0.85***	0.81	0.82	0.82
		Avg. (Std.)	**0.88 (0.015)**	0.77 (0.061)	0.78 (0.059)	0.80 (0.073)	0.79 (0.052)	0.80 (0.048)	***0.81 (0.046)***
		Test time	**~148ms**	~2.1s	~430ms	***~259.08s***	~4.03s	~7.06s	~2.86s
DDTI	Adam	Dice Coeff.	**0.85**	0.78	0.80	***0.82***	0.80	***0.82***	0.81
		IoU	**0.81**	0.65	0.68	0.72	0.68	0.70	***0.73***
		Precision	**0.88**	0.80	0.80	***0.82***	0.80	***0.82***	***0.82***
		Recall	**0.88**	0.75	0.77	0.75	0.77	***0.80***	0.78
		Avg. (Std.)	**0.86 (0.029)**	0.75 (0.058)	0.76 (0.049)	0.78 (0.044)	0.76 (0.049)	0.78 (0.049)	***0.79 (0.035)***

Dataset	Optimizer	Metrics	LASegNet (Proposed)	U-Net [21]	Attention U-Net [87]	Dense U-Net [88]	HyperDense-Net [89]	TransUNet [90]	MISSFormer [91]
		No. of parameters	**2.2 M**	31.04 M	6.4 M	***4 M***	60 M	105 M	42.46 M
	fAdamR	Dice Coeff.	**0.87**	0.80	0.82	0.83	0.82	***0.84***	***0.84***
		IoU	**0.80**	0.67	0.70	0.73	0.71	0.73	***0.74***
		Precision	**0.89**	0.82	0.83	0.84	0.83	***0.85***	***0.85***
		Recall	**0.90**	0.78	0.79	0.79	0.80	0.82	***0.83***
		Avg. (Std.)	**0.87 (0.039)**	0.76 (0.058)	0.78 (0.051)	0.79 (0.043)	0.79 (0.047)	0.81 (0.047)	***0.82 (0.041)***
		Test time	**~49ms**	690ms	142ms	***87ms***	1.3s	2.2s	943ms
Overall avg. for Adam (Std.)			**0.87 (0.026)**	0.75 (0.054)	0.77 (0.050)	0.78 (0.055)	0.78 (0.052)	0.80 (0.049)	***0.81 (0.035)***
Overall avg. for fAdamR (Std.)			**0.89 (0.023)**	0.77 (0.056)	0.79 (0.053)	0.80 (0.058)	0.80 (0.049)	0.82 (0.047)	***0.84 (0.032)***

Red: first best, ***blue with italic: second best*** and green: third best

Visual Analysis

Figure 4.3 provides a visual illustration of the segmentation performance of the LASegNet (proposed), MISSFormer, and TransUNet. Note that, for this comparison, we have chosen the latter to segmentation schemes since they provide, in general, the second best and third best qualitative performance, respectively. For this purpose, we have selected one typical image from each of the six datasets shown in the first column of the figure. The second, the third, the fourth, and the fifth columns of this figure illustrate, respectively, the corresponding ground truth segmentation, and segmentations resulting from the proposed LASegNet, MISSFormer, and TransUNet. By comparing the results, especially those of each of the two boxes, that the proposed network is able to provide a better segmentation.

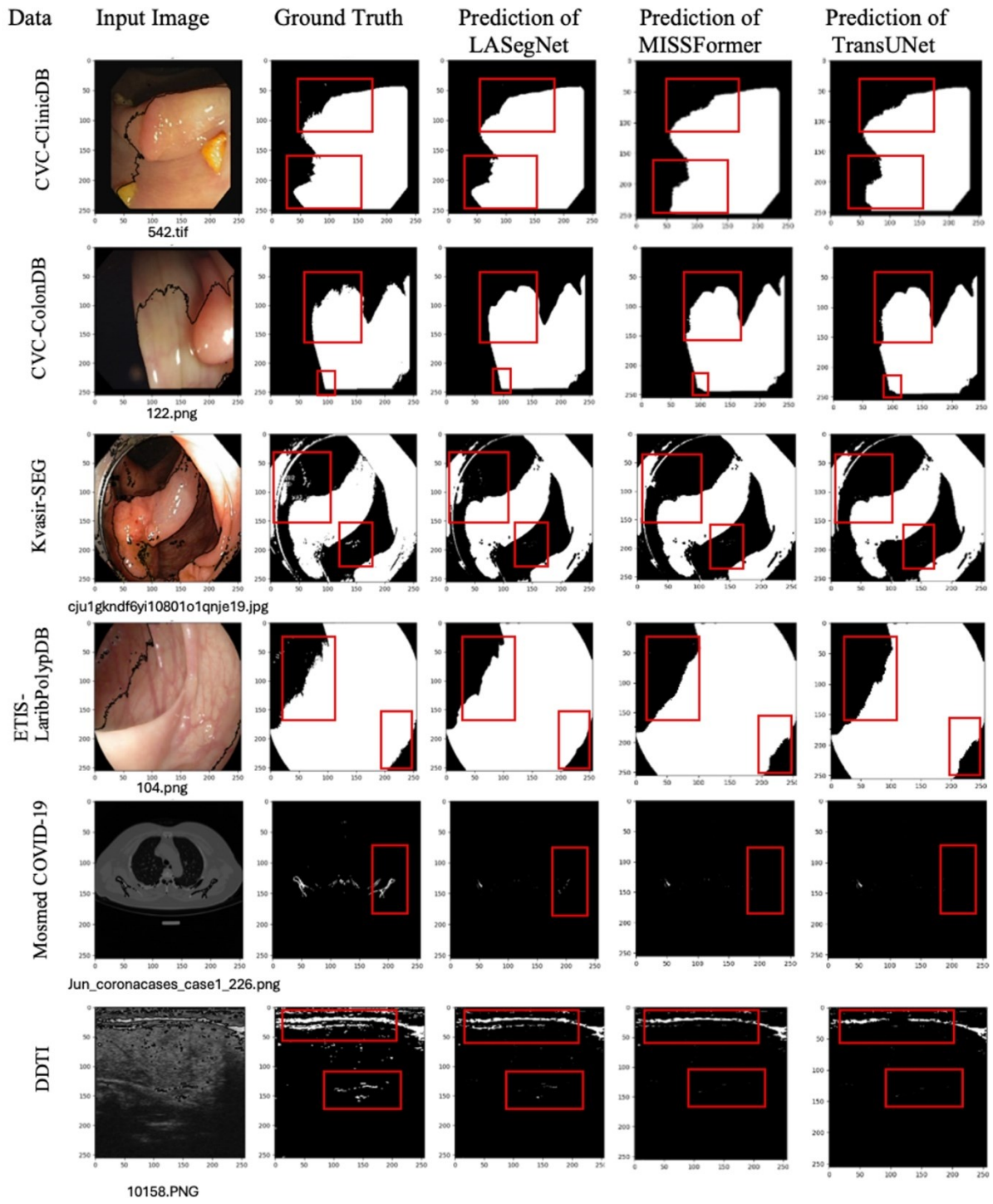


Figure 4.3: Visual illustration of the segmentation performance of the proposed LASegNet network, MISSFormer, and TransUNet

4.7 Summary

In this chapter, a novel light-weight attention-guided segmentation network with feature recalibration, LASegNet, has been proposed for segmenting various types of medical images. The proposed network has been designed as a light-weight variant of the U-Net by significantly reducing the number of filters in each block of the encoder and the decoder without compromising the performance by adding squeeze-and-excitation (SE) blocks and attention modules. Our network has a series of encoding blocks for feature extraction, a series of decoding blocks for feature reconstruction, squeeze-and-excitation (SE) blocks for feature recalibration, a central block for bridging the encoder and the decoder, and attention modules for paying more attention to the important features. The proposed scheme has been applied on six different datasets. It has been seen that the number of parameters used by the proposed network is significantly smaller than that of the state-of-the-art networks with a performance better than that of latter networks. The experimental results have shown that the proposed scheme provides a performance on images from the different modalities with a largest average value and a smallest standard deviation compared to that of the other schemes, indicating robustness of the proposed scheme across different modalities of medical images.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

Medical image segmentation is the process of isolating the object of interest from medical images of a given modality to enable quantitative and qualitative analysis for supporting accurate diagnosis, treatment planning, and disease monitoring. That object of interest could be an entire organ or a certain region of abnormalities within the organ. In this thesis, we have proposed five different schemes for medical image segmentation. We have proposed four different schemes in Chapter 3, MAGNet, MedSegNet, SSNet, and FFNet, utilizing attention mechanism-enhanced multiscale feature fusion networks, where each network has been designed to address specific challenges across different imaging modalities. The networks, MAGNet and MedSegNet, have been proposed for segmenting CT, colonoscopy, and non-mydriatic 3CCD images, with MedSegNet offering a lighter weight architecture with a performance comparable to that of the MAGNet. Further, SSNet has been proposed to effectively utilize both labeled and unlabeled data for segmenting the brain anatomical structures of tissues in MRI images, particularly in cases where labeled data are significantly fewer than unlabeled data. In addition, FFNet, a feature fusion-based network, has been proposed for segmenting breast cancer images, employing a convolutional neural network with a multipath encoder. It has been shown that these four schemes exhibit a

performance superior to that of the existing state-of-the-art networks in terms of the standard performance metrics. However, these four networks require a large number of parameters. To address this problem without compromising the performance, we have proposed in Chapter 4 a lightweight network with feature recalibration and attention mechanisms for segmenting medical images. This network consists of a series of encoding blocks for feature extraction, a series of decoding blocks for feature reconstruction, squeeze-and-excitation (SE) blocks for feature recalibration, a central block for bridging the encoder and the decoder, and attention modules for paying more attention to the important features. We have significantly reduced the number of filters in each block of the encoder and the decoder to reduce the complexity of the network. This reduction in the number of filters will affect the segmentation performance of the network. To compensate for the deterioration of the performance, the squeeze-and-excitation (SE) and the attention modules are introduced. However, the complexity overhead caused by the addition of these blocks is only minimal. Extensive experiments have been carried out on six different datasets for segmenting polyp from endoscopy and coloscopy, lung infection from CT scan and thyroid from ultrasound images. It has been shown that the parameters used by the proposed network is significantly lower than that of the existing state-of-the-art networks, while at the same time its performance being better than that of the other networks, irrespective of the modality. Experiments have also shown that the proposed scheme provides a segmentation performance with the largest average value and the smallest standard deviation for the performance metrics, indicating the robustness of the proposed lightweight network in segmenting images of different modalities.

5.2 Scope for Future Work

Based on the work presented in this thesis, further work can be carried out to explore the applicability of the proposed LASegNet network to any other modalities. One limitation of the

current network is its reliance on fully annotated data. To overcome the data annotation problem, unsupervised and self-supervised learning strategies can be studied on unlabeled data for reducing the annotation costs; this is particularly useful for rare diseases for which suitable datasets are not available, let alone annotated datasets. Further, domain adaptation techniques can be utilized to improve generalization across heterogeneous datasets. Incorporation of vision-language models may also provide a deeper contextual understanding and open new directions for multi-modal medical image segmentation.

References

[1] R. Wang, T. Lei, R. Cui, B. Zhang, H. Meng, and A. K. Nandi, "Medical image segmentation using deep learning: A survey," *IET Image Process.*, vol. 16, no. 5, pp. 1243–1267, 2022.

[2] L. Zhang *et al.*, "Block level skip connections across cascaded V-Net for multi-organ segmentation," *IEEE Trans. Med. Imaging*, vol. 39, no. 9, pp. 2782–2793, 2020.

[3] L. Liu, J. M. Wolterink, C. Brune, and R. N. Veldhuis, "Anatomy-aided deep learning for medical image segmentation: a review," *Phys. Med. Biol.*, vol. 66, no. 11, p. 1-22, 2021.

[4] Z. Pan and J. Lu, "A Bayes-based region-growing algorithm for medical image segmentation," *Comput. Sci. Eng.*, vol. 9, no. 4, pp. 32–38, 2007.

[5] J. Yang and S.-C. Huang, "Method for evaluation of different MRI segmentation approaches," *IEEE Trans. Nucl. Sci.*, vol. 46, no. 6, pp. 2259–2265, 1999.

[6] P.-H. Conze, G. Andrade-Miranda, V. K. Singh, V. Jaouen, and D. Visvikis, "Current and emerging trends in medical image segmentation with deep learning," *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 7, no. 6, pp. 545–569, 2023.

[7] D.-P. Fan *et al.*, "Inf-net: Automatic covid-19 lung infection segmentation from ct images," *IEEE Trans. Med. Imaging*, vol. 39, no. 8, pp. 2626–2637, 2020.

[8] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," in *Proc. 27th Int. Joint Conf. on Artif. Intell.*, Stockholm, Sweden, 2018, pp. 698-704.

[9] H. Polat, "A modified DeepLabV3+ based semantic segmentation of chest computed tomography images for COVID-19 lung infections," *Int. J. Imaging Syst. Technol.*, vol. 32, no. 5, pp. 1481–1495, 2022.

[10] J. Ma *et al.*, "Towards efficient COVID-19 CT annotation: A benchmark for lung and infection segmentation," *Medical Physics*, vol. 48, no. 3, pp. 1197-1210, 2021.

[11] D. Jha *et al.*, "Real-time polyp detection, localization and segmentation in colonoscopy using deep learning," *IEEE Access*, vol. 9, pp. 40496–40510, 2021.

[12] D.-P. Fan *et al.*, "Pranet: Parallel reverse attention network for polyp segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2020, pp. 263–273.

[13] C.-H. Huang, H.-Y. Wu, and Y.-L. Lin, "Hardnet-mseg: A simple encoder-decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 fps," *ArXiv Prepr. ArXiv210107172*, 2021.

[14] L. Mou *et al.*, "CS-Net: Channel and spatial attention network for curvilinear structure segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2019, pp. 721–730.

[15] Y. Zhou, H. Yu, and H. Shi, "Study group learning: Improving retinal vessel segmentation trained with noisy labels," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.,* Springer, 2021, pp. 57–67.

[16] S. A. Kamran, K. F. Hossain, A. Tavakkoli, S. L. Zuckerbrod, K. M. Sanders, and S. A. Baker, "RV-GAN: Segmenting retinal vascular structure in fundus photographs using a novel multi-scale generative adversarial network," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Intervention*, Springer, 2021, pp. 34–44.

[17] J. Schlemper *et al.*, "Attention gated networks: Learning to leverage salient regions in medical images," *Med. Image Anal.*, vol. 53, pp. 197–207, 2019.

[18] W. Liu *et al.*, "Full-resolution network and dual-threshold iteration for retinal vessel and coronary angiograph segmentation," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 9, pp. 4623–4634, 2022.

[19] N. K. Tomar et al., "Fanet: A feedback attention network for improved biomedical image segmentation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 11, pp. 9375 - 9388, 2022.

[20] J. Schlemper *et al.*, "Attention gated networks: Learning to leverage salient regions in medical images," *Med. Image Anal.*, vol. 53, pp. 197–207, 2019.

[21] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Interv.,* Springer, 2015, pp. 234–241.

[22] S. Bharati, M. O. Ahmad, and M.N.S. Swamy, "A Novel Continual Learning Approach for Robust Medical Image Segmentation," in *Proc. 2025 IEEE 68th Int. Midwest Symp. on*

Circuits and Systems (MWSCAS 2025), Lansing, Michigan, USA, IEEE, Aug. 2025, pp. 965-969.

[23] A. Vaswani, "Attention is all you need," in *Proc. 31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, 2017, pp. 1-11.

[24] M. Yi-de, L. Qing, and Q. Zhi-Bai, "Automated image segmentation using improved PCNN model based on cross-entropy," in *Proc. 2004 Int. Symp. on Intell. Multi., Video and Speech Process.*, 2004, IEEE, 2004, pp. 743–746.

[25] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham, Switzerland, Springer, 2017, pp. 240–248.

[26] D. P. Kingma, J. Ba, "Adam: A method for stochastic optimization," *Int. Conf. on Learn. Rep. (ICLR)*, 2015, pp. 1-13.

[27] I. Loshchilov and F. Hutter, "Fixing weight decay regularization in adam," *ArXiv Prepr. ArXiv171105101*, vol. 5, 2017.

[28] F. Zou, L. Shen, Z. Jie, W. Zhang, and W. Liu, "A sufficient condition for convergences of adam and rmsprop," in *Proc. IEEE/CVF Conf. on comp. vision and pattern recog.*, 2019, pp. 11127–11135.

[29] G. Hinton, N. Srivastava, and K. Swersky, "Neural networks for machine learning lecture 6a overview of mini-batch gradient descent," Dept. Comput. Sci., Toronto Univ., Toronto, ON, Canada, Tech. Rep., Jul. 2012.

[30] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proc. 19th Int. Conf. Comput. Statist.*, 2010, pp. 177–186.

[31] R. Ward, X. Wu, and L. Bottou, "Adagrad stepsizes: Sharp convergence over nonconvex landscapes," *J. Mach. Learn. Res.*, vol. 21, no. 219, pp. 1–30, 2020.

[32] J. Yu *et al.*, "Learning generalized intersection over union for dense pixelwise prediction," in *Proc. Int. Conf. on Mach. Learn.*, PMLR, 2021, pp. 12198–12207.

[33] T. Eelbode *et al.*, "Optimization for medical image segmentation: theory and practice when evaluating with dice score or jaccard index," *IEEE Trans. Med. Imaging*, vol. 39, no. 11, pp. 3679–3690, 2020.

[34] D. Jha *et al.*, "Resunet++: An advanced architecture for medical image segmentation," in *Proc. 2019 IEEE Int. Symp. on Multi. (ISM)*, IEEE, 2019, pp. 225–2255.

[35] S. Bharati, P. Podder, and M. R. H. Mondal, "Hybrid deep learning for detecting lung diseases from X-ray images," *Inform. Med. Unlocked*, vol. 20, p. 100391, 2020.

[36] S. Bharati, M. O. Ahmad, and M.N.S. Swamy, "FewShotEEG Learning and Classification for Brain-Computer Interface," in *Proc. 2024 IEEE Int. Symp. on Circuits and Systems (ISCAS)*, Singapore, IEEE, 2024, pp. 1–5.

[37] S. Bharati, M. O. Ahmad, and M.N.S. Swamy, "MAGNet: A Convolutional Neural Network with Multi-Scale and Global Attention Modules for Medical Image Segmentation," in *Proc. 2024 IEEE Int. Symp. on Circuits and Systems (ISCAS)*, Singapore, IEEE, 2024, pp. 1–5.

[38] S. Bharati, M. O. Ahmad, and M.N.S. Swamy, "MedSegNet: A Lightweight Convolutional Network Combining Dual Self-Attention and Multi-Scale Attention for Medical Image Segmentation," in *Proc. 2024 IEEE 67th Int. Midwest Symp. on Circuits and Systems (MWSCAS)*, Springfield, MA, US, IEEE, 2024, pp. 965–969.

[39] S. Bharati, M. O. Ahmad, and M.N.S. Swamy, "Dual Task Learning: A Semi-Supervised Approach to Medical Image Joint Segmentation and Registration," in *Proc. 2025 IEEE Int. Symp. on Circuits and Systems (ISCAS 2025)*, London, UK, IEEE, 2025, pp. 1–5.

[40] S. Bharati, M. O. Ahmad, and M.N.S. Swamy, "A Novel Super-pixel Grid Mixing-Based Augmentation with a Feature Fusion of Convolutional Networks for Breast Ultrasound Image Segmentation," in *Proc. 2024 IEEE 67th Int. Midwest Symp. on Circuits and Systems (MWSCAS 2024)*, Springfield, MA, US, IEEE, 2024, pp. 970–974.

[41] A. Sinha and J. Dolz, "Multi-scale self-guided attention for medical image segmentation," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 1, pp. 121–130, 2020.

[42] J. Liu, Q. Chen, Y. Zhang, Z. Wang, X. Deng, and J. Wang, "Multi-level feature fusion network combining attention mechanisms for polyp segmentation," *Inf. Fusion*, vol. 104, p. 102195, 2024.

[43] https://medicalsegmentation.com/covid19/ (Accessed Oct. 24, 2023).

[44] https://polyp.grand-challenge.org/CVCClinicDB/ (Accessed Oct. 24, 2023).

[45] https://drive.grand-challenge.org/ (Accessed Oct. 24, 2023).

[46] A. Galassi, M. Lippi, and P. Torroni, "Attention in natural language processing," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 10, pp. 4291–4308, Oct. 2021.

[47] S. Huang, D. Wang, X. Wu, and A. Tang, "Dsanet: Dual self-attention network for multivariate time series forecasting," in *Proc. 28th ACM Int. Conf. Inf. Knowl. Manage.*, 2019, pp. 2129–2132.

[48] Y. Huang, W. Liu, C. Li, Y. Liang, H. Yang, and F. Meng, "MSANet: A multi-scale attention module," in *Proc. Conf. Intell. Syst. Knowl. Eng.*, Dalian, China, 2019, pp. 659–663.

[49] M. Liao, H. Tang, X. Li, P. Vijayakumar, V. Arya, and B. B. Gupta, "A lightweight network for abdominal multi-organ segmentation based on multi-scale context fusion and dual self-attention," *Inf. Fusion*, vol. 108, p. 102401, 2024.

[50] L. F. Sánchez-Peralta, A. Picón, F. M. Sánchez-Margallo, and J. B. Pagador, "Unravelling the effect of data augmentation transformations in polyp segmentation," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 15, no. 12, pp. 1975–1988, 2020.

[51] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: fast and flexible image augmentations," *Information*, vol. 11, no. 2, p. 125, 2020.

[52] "OASIS-1." [Online]. Available: https://sites.wustl.edu/oasisbrains/ (Accessed Oct. 18, 2024).

[53] Z. Xu and M. Niethammer, "DeepAtlas: Joint semi-supervised learning of image registration and segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Shenzhen, China: Springer, 2019, pp. 420–429.

[54] M. J. Cardoso *et al.*, "Monai: An open-source framework for deep learning in healthcare," *ArXiv Prepr. ArXiv221102701*, 2022.

[55] A. Fajar, R. Sarno, C. Fatichah, and A. Fahmi, "Reconstructing and resizing 3D images from DICOM files," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 34, no. 6, pp. 3517–3526, 2022.

[56] R. Zhao *et al.*, "Rethinking dice loss for medical image segmentation," in *Proc. 2020 IEEE Int. Conf. on Data Mining (ICDM)*, IEEE, 2020, pp. 851–860.

[57] M. L. Terpstra, M. Maspero, A. Sbrizzi, and C. A. van den Berg, "⊥-loss: A symmetric loss function for magnetic resonance imaging reconstruction and image registration with deep learning," *Med. Image Anal.*, vol. 80, p. 102509, 2022.

[58] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, 2017.

[59] V. Badrinarayanan, A. Kendall, and R. Cipolla, “Segnet: A deep convolutional encoder-decoder architecture for image segmentation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, 2017.

[60] C. Szegedy *et al.*, “Going deeper with convolutions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[61] N. Yamanakkanavar and B. Lee, “MF2-Net: A multipath feature fusion network for medical image segmentation,” *Eng. Appl. Artif. Intell.*, vol. 114, p. 105004, 2022.

[62] A. Iqbal and M. Sharif, “Memory-efficient transformer network with feature fusion for breast tumor segmentation and classification task,” *Eng. Appl. Artif. Intell.*, vol. 127, p. 107292, 2024.

[63] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, “Dataset of breast ultrasound images,” *Data Brief*, vol. 28, p. 104863, 2020.

[64] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, “Cutmix: Regularization strategy to train strong classifiers with localizable features,” in *Proc. IEEE/CVF Int. Conf. on Comp. vision*, IEEE, 2019, pp. 6023–6032.

[65] C. Thomas, M. Byra, R. Marti, M. H. Yap, and R. Zwiggelaar, “BUS-Set: A benchmark for quantitative evaluation of breast ultrasound segmentation networks with public datasets,” *Med. Phys.*, vol. 50, no. 5, pp. 3223–3243, 2023.

[66] Z. Zhuang, N. Li, A. N. Joseph Raj, V. G. Mahesh, and S. Qiu, “An RDAU-NET model for lesion segmentation in breast ultrasound images,” *PloS One*, vol. 14, no. 8, p. e0221535, 2019.

[67] L. Han *et al.*, “Semi-supervised segmentation of lesion from breast ultrasound images with attentional generative adversarial network,” *Comput. Methods Programs Biomed.*, vol. 189, p. 105275, 2020.

[68] A. Vakanski, M. Xian, and P. E. Freer, “Attention-enriched deep learning model for breast tumor segmentation in ultrasound images,” *Ultrasound Med. Biol.*, vol. 46, no. 10, pp. 2819–2833, 2020.

[69] A. A. Hekal, A. Elnakib, H. E.-D. Moustafa, and H. M. Amer, “Breast cancer segmentation from ultrasound images using deep dual-decoder technology with attention network,” *IEEE Access*, vol. 12, pp. 10087–10101, 2024.

[70] N. S. Punn and S. Agarwal, "RCA-IUnet: a residual cross-spatial attention-guided inception U-Net model for tumor segmentation in breast ultrasound imaging," *Mach. Vis. Appl.*, vol. 33, no. 2, p. 27, 2022.

[71] Y. Lyu, Y. Xu, X. Jiang, J. Liu, X. Zhao, and X. Zhu, "AMS-PAN: Breast ultrasound image segmentation model combining attention mechanism and multi-scale features," *Biomed. Signal Process. Control*, vol. 81, p. 104425, 2023.

[72] T. Chavan, K. Prajapati, and K. R. JV, "InvUNET: Involuted UNET for breast tumor segmentation from ultrasound," in *Proc. 20th Int. Conf. Artif. Intell. Med. (AIME)*. Halifax, NS, Canada: Springer, Jun. 2022, pp. 283–290.

[73] S. Bharati, M. O. Ahmad, and M.N.S. Swamy, "A Novel Lightweight Attention-Guided Network with Feature Recalibration for Medical Image Segmentation," *under review for journal publication*, 2025.

[74] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. on Comp. vision and Patt. Recog.*, IEEE, 2018, pp. 7132–7141.

[75] J. Zhang, Z. Jiang, J. Dong, Y. Hou, and B. Liu, "Attention gate resU-Net for automatic MRI brain tumor segmentation," *IEEE Access*, vol. 8, pp. 58533–58545, 2020.

[76] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Comput. Med. Imaging Graph.*, vol. 43, pp. 99–111, 2015.

[77] J. Bernal, J. Sánchez, and F. Vilarino, "Towards automatic polyp detection with a polyp appearance model," *Pattern Recognit.*, vol. 45, no. 9, pp. 3166–3182, 2012.

[78] D. Jha *et al.*, "Kvasir-seg: A segmented polyp dataset," in *Proc. Int. Conf. Multimedia Modeling*, 2020, pp. 451–462.

[79] J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, "Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 9, pp. 283–293, 2014.

[80] S. P. Morozov *et al.*, "Mosmeddata: Chest ct scans with covid-19 related findings dataset," *ArXiv Prepr. ArXiv200506465*, 2020.

[81] L. Pedraza, C. Vargas, F. Narváez, O. Durán, E. Muñoz, and E. Romero, "An open access thyroid ultrasound image database," in *Proc. 10th Int. Symp. on Med. Infor. Process. and Anal.,* SPIE, 2015, pp. 188–193.

[82] G. Wang *et al.*, "DeepIGeoS: a deep interactive geodesic framework for medical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1559–1572, 2018.

[83] J. Xiao, W. Guo, and J. Liu, "Exploring Data Augmentation Effects on A Singular Illumination Distribution Dataset with ColorJitter," in *Proc. 2024 3rd Int. Conf. on Image Process. and Media Comput. (ICIPMC),* IEEE, 2024, pp. 75–81.

[84] X. Li, L. Yu, H. Chen, C.-W. Fu, L. Xing, and P.-A. Heng, "Transformation-consistent self-ensembling model for semisupervised medical image segmentation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 523–534, 2020.

[85] *TensorFlow*. [Online]. Available: https://www.tensorflow.org/ (Accessed Oct. 18, 2024).

[86] *Keras*. [Online]. Available: https://keras.io/ (Accessed Oct. 18, 2024).

[87] O. Oktay *et al.*, "Attention u-net: Learning where to look for the pancreas," *1st Conf. on Med. Imaging with Deep Learning (MIDL 2018)*, Amsterdam, The Netherlands, 2018, pp. 1-10.

[88] S. Cai, Y. Tian, H. Lui, H. Zeng, Y. Wu, and G. Chen, "Dense-UNet: a novel multiphoton in vivo cellular image segmentation model based on a convolutional neural network," *Quant. Imaging Med. Surg.*, vol. 10, no. 6, p. 1275, 2020.

[89] J. Dolz, K. Gopinath, J. Yuan, H. Lombaert, C. Desrosiers, and I. B. Ayed, "HyperDense-Net: a hyper-densely connected CNN for multi-modal image segmentation," *IEEE Trans. Med. Imaging*, vol. 38, no. 5, pp. 1116–1126, 2018.

[90] J. Chen *et al.*, "TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers," *Med. Image Anal.*, vol. 97, p. 103280, 2024.

[91] X. Huang, Z. Deng, D. Li, X. Yuan, and Y. Fu, "Missformer: An effective transformer for 2d medical image segmentation," *IEEE Trans. Med. Imaging*, vol. 42, no. 5, pp. 1484–1494, 2022.