Exploring Resource-Efficient Deep Learning for Medical Image Segmentation

A thesis submitted to Indian Statistical Institute
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computer Science

by

Pallabi Dutta
Senior Research Fellow

Under the supervision of

Prof. Sushmita Mitra

Machine Intelligence Unit
Indian Statistical Institute,
Kolkata, India

May, 2026

Certificate from Supervisor

This is to certify that the work contained in the thesis entitled **"Exploring Resource-Efficient Deep Learning for Medical Image Segmentation"**, submitted by **Pallabi Dutta** for the award of the degree of Doctor of Philosophy in Computer Science to Indian Statistical Institute, Kolkata, is a record of the bonafide research works carried out by her under my supervision and guidance.

I consider that the thesis has reached the standards and fulfilling the requirements of the rules and regulations relating to the nature of the degree. The contents embodied in the thesis have not been submitted as a whole or as a part for the award of any degree or diploma or any other academic award anywhere given before.

Signature: Sushmita Mitra

Date: 18.05.2026

Prof. Sushmita Mitra
Professor, Machine Intelligence Unit,
Indian Statistical Institute, Kolkata,
203 Barrackpore Trunk Road,
Kolkata, West Bengal , India - 700108

Dedicated to friends and family...

Acknowledgements

Writing this thesis has been a journey of immense learning and personal growth. It would not have been possible without the support of many individuals.

First, I would like to extend my gratitude to my supervisor, Prof. Sushmita Mitra, whose mentorship has been crucial to my academic and professional growth. Her valuable suggestions and keen attention to detail have been invaluable in shaping this thesis. I am deeply grateful for the time she invested in reviewing my work. I am thankful for the academic freedom she granted me to explore my ideas.

I am grateful to my lab mates, Dr. Surochita Pal, Shramana Dey, Riddhasree Bhattacharyya, and Anubhab Maity, for their support and suggestions during the experimental phase of my research. I would also like to thank Prof. B. Uma Shankar, Dr. Swalpa Kumar Roy, and Soham Bose for their constructive feedback. I sincerely thank the Machine Intelligence Unit and IDEAS-TIH, ISI Kolkata, for granting me access to their computing facilities, required to execute my experiments. I also extend my thanks to my friends, Ankita Dutta, Sayan Saha, Chayan Maitra, and Soumi Sarkar for providing a cheerful atmosphere.

Most importantly, I express my deepest gratitude to my parents. My father often reminds me of the importance of taking a difficult path, stating his favourite lines, " *I took the one less traveled by, and that has made all the difference*". This journey was indeed a challenging road; however, looking back, it has made all the difference in who I have become today. I thank my husband for being the pillar of strength in my tough times. Your patience and understanding gave me the courage to continue.

Pallabi Dutta

Abstract

Automated medical image segmentation improves diagnostic accuracy by automating the precise delineation of target anatomical structures in the input images. Artificial Intelligence (AI), and specifically, Deep Learning (DL), has emerged as a state-of-the-art approach for this task. However, the significant computational demands of DL approaches often hinders their deployment. Advanced models, including Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), require substantial processing power and a large memory footprint, limiting their use in resource-constrained settings. This thesis aims to address this challenge by developing a series of novel, resource-efficient DL models that achieve high segmentation accuracy with reduced computational costs. The research follows a logical progression of architectural novelty. First, global context-aware attention frameworks, *FuDSA-Net* and *VoCANet*, are introduced by leveraging multi-scalar features and global-context aware attention for efficient 2D/3D segmentation. The spatial and spectral domains are then integrated using a novel hybrid CNN-ViT framework *WaveCoformer* for learning robust representation of the target structure. The developed model achieves high segmentation accuracy with a lower parameter count. Subsequently, the research investigates a computationally efficient alternative to ViTs for segmentation, called Vision-xLSTM, by developing the *U-VixLSTM* model. This is extended to the *Rot-UViL* architecture, capable of modeling cross-dimensional dependencies in volumetric inputs with its novel rotational attention. Finally, the thesis presents a prompt-driven pruning framework for ViT-based segmentation models, called *PrATo*, which dynamically prunes irrelevant ViT tokens with a parameter-free prompt-driven scoring mechanism. The framework achieves $\sim 35 - 55\%$ reduction of processed tokens. The frameworks developed in this thesis are validated across multiple publicly available datasets; demonstrating their high segmentation accuracy along with computational efficiency.

Contents

List of Figures

List of Tables

List of Abbreviations

DSC	**D**ice **S**imilarity **C**oefficient
Precision	**Precision** Metric
Recall/Sensitivity	**Recall** Metric
*HD*95	**H**ausdorff **D**istance **95**%
Accuracy	**Accuracy** Metric
*F*1	**F1**-Score
ROI	**R**egion **O**f **I**nterest
TP	**T**rue **P**ositive
TN	**T**rue **N**egative
FP	**F**alse **P**ositive
FP	**F**alse **N**egative
DL	**D**ice **L**oss
CE	**C**ross **E**ntropy
TL	**T**versky **L**oss
FL	**F**ocal **L**oss
FTL	**F**ocal **T**versky **L**oss
HU	**H**ounsfield **U**nit

List of Symbols

I	: 2-dimensional Image
$\forall$	: for all
S	: set of n sub-regions of I
ϕ	: null set
$\cup$	: Union
$\cap$	: Intersection
$\neq$	: Not equal to
$\wedge$	: Logical AND
$\implies$	: Implies
$\psi(.)$	: wavelet function
$\alpha(.)$	: scaling function
l	: Model stages
A	: Attention mechanism
C, H, W, D	: Channel, Height, Width and Depth dimensions
λ	: 2D atrous (dilated) convolution
δ	: 3D atrous (dilated) convolution
r	: Dilation rate
GAP	: Global Average Pooling
σ	: Sigmoid activation
β	: Bilinear upsampling
τ	: Trilinear upsampling
$\mathbf{L}$	: Loss function
Ω	: Model weights
Γ	: Ground truth
$\hat{\Gamma}$	: Prediction map
MLP	: Multi Layer Perceptron
DWT	: Discrete Wavelet Transform
X_{HF}	: High-frequency component
X_{LF}	: Low-frequency component
$w_{x\times x\times x}$	: $x \times x \times x$ 3D convolution
θ	: Window Multi-head self-attention
θ'	: Shifted Window Multi-head self-attention
λ	: LeakyReLU activation
$\oplus$	: Element-wise addition
$\otimes$	: Element-wise multiplication
Q, K, V	: Query, Key and Value vectors
η	: Softmax activation
P	: Patch dimension
K_{pos}	: Position embedding matrix
$Concat(.)$	: Feature Map Concatenation

Chapter 1

Introduction and Scope of Thesis

1.1 Background

Medical imaging [122] is a non-invasive method for visualizing human anatomy. It is an essential tool for identifying abnormalities such as tumor, lesion, infection, inflammation, and blockage. This helps medical professionals predict the stage or severity of a disease by analyzing the size, location, and extent of the abnormality in the body. Recent innovations in healthcare, such as interventional radiology [107], utilize medical imaging for non-invasive or minimally invasive procedures in the diagnosis and treatment of a wide range of diseases. Medical imaging also has a significant contribution to the advancement of medical research. Imaging combined with genetic information and symptoms allows a better understanding of the characteristics of any disease [17]. This leads to new strategies such as personalized medicine [90].

Different imaging modalities, each with their unique characteristics, facilitate visualization of human anatomy and disease progression; thereby, offering useful clinical insights. Computed Tomography (CT) scans use X-rays to generate cross-sectional images of the target anatomical structure. It may or may not involve the administration of contrast agents in humans. This enhances the visualization of the target structure(s) with clarity. It is suitable for diagnosing a wide range of conditions, including tumors, fractures, and cardiovascular and cerebral factors. However, the contrast agent can also induce allergic reactions and renal complications [49]. Magnetic Resonance Imaging (MRI) employs powerful magnetic fields to generate intricate images of specific areas within the human body. They are superior in detail resolution, making them optimal for diagnosing diseases related to the brain and spinal cord. Significant imaging expenses, long scanning durations, and inappropriateness for those with implants are some of the drawbacks associated with MRIs.

Optical imaging modalities such as fundus imaging and dermoscopy provide crucial information on diseases related to the eyes and skin, respectively. Fundus imaging focuses on the inner wall of the back of the eye and is used by ophthalmologists to diagnose diseases, including diabetic retinopathy, macular degeneration, and glaucoma. Imaging data helps to detect and study the progression of associated lesions, such as hemorrhage, microaneurysms and hard and soft exudates. Dermoscopy uses optical magnification and a special lighting setup to diagnose skin diseases. It helps in detecting skin cancer, seborrheic keratose, etc.

Analyzing the diverse nature of clinical insights from these different types of imaging modalities requires advanced analytical methods and functionalities. Medical image segmentation is a crucial tool for delineating target anatomical structures from medical images. It divides an image into segments, with each partition representing some anatomical structure or disease region [7], [114]. Subsequently essential biomarkers, *viz.* volume, shape and size of lesions, tissues, or organs, may be extracted from the delineated Region of Interest (ROI). These are needed by clinicians to estimate the prognosis of a disease. Radiotherapists quantify radiation dose in cancer by analyzing biomarkers to determine the extent of the disease and its possible risk to healthy tissues [126]. The

precise delineation of the voxels, associated with the target anatomical components in the volumetric images, facilitates their reconstruction in interactive 3D models [5], [100]. Clinicians obtain a thorough understanding of the actual geometry of the target anatomy by examining the 3D model from various perspectives and scales. Virtual surgical planning [135], designing implants and prosthetics, providing advanced medical education and training are some of the highly impactful applications of these 3D models.

Segmentation of target organs or tissues can be performed manually, semi-automatically, or automatically. Manual segmentation requires a precise delineation by specialists of the relevant structures, slice-by-slice, from volumetric images such as CT scans and MRIs. The task requires extensive attention to detail over an extended time span and can be exhausting for the human expert involved. Therefore, it can cause inconsistencies in the final output, leading to inter- and/or intra-rater variability. Huge expenses and the availability of trained professionals can be a potential bottleneck in resource-constrained environments. The reconstructed anatomical structure, obtained by stacking multiple manually segmented slices from volumetric images, might not have a smooth surface like real biological structures. This results in discrepancies when interpreting the final 3D virtual representation of the target areas. The constraints of time, consistency, and cost emphasize the necessity for automated techniques that can efficiently segment medical images [10].

Semi-automated segmentation was developed based on human-computer interaction to reduce the burden associated with purely manual segmentation techniques [99], [13], [44]. These algorithms were assisted by cues provided by human experts. Some common categories of semi-automated segmentation are as follows.

- *Region growing [2]:* The user initially selects one or more seed pixels/voxels in the input image. The algorithm recursively expands the regions around the seed pixel(s) by including the neighboring pixels/voxels based on similarity measures. They are advantageous for segmenting structures with homogeneous surfaces.

- *Active contours [66]:* An initial contour is approximated by user around a target structure with binary array. The algorithm iteratively refines the contour to better fit the target structure, using elastic energy (for smoothness), curvature analysis (for preventing sharp bends), and gradient calculations (for attracting contours towards the ROIs).

- *Graph cut [14]:* Each pixel/voxel in the input image acts as a node in the graph. An edge between any two nodes represents the similarity between the pixels with respect to intensity level/brightness. Every edge is associated with weights that indicate the likelihood that a pixel belongs to the foreground or background regions. The algorithm computes the minimum cut in the corresponding graph that groups similar pixels together; thereby, separating different regions in the given image.

Although semi-automated approaches reduced the burden of segmentation tasks compared to fully manual approaches, they still required a considerable

amount of domain knowledge. For example, the success of the region-growing and active-contour approaches depends on the initial placement of the seed(s). This makes the process subjective to the user's knowledge and, therefore, prone to error in clinical scenarios where precision is of utmost importance. Such persistent challenges necessitated the development of automated segmentation strategies.

Artificial Intelligence (AI) refers to computer systems that learn to solve tasks requiring human intelligence [105]. A significant chunk of AI is Machine Learning (ML), which is a group of statistical algorithms to discover hidden patterns from input data without any rule-based algorithm [89]. Deep Learning (DL) [69] is a subset of ML that is conceptually inspired by Artificial Neural Networks (ANNs); particularly Multi-Layer Perceptrons (MLPs) [110] which stack multiple layers of perceptrons to effectively learn non-linear functions. Although ANNs constituted shallow ML, the deeper architectures of DL revolutionized their quest for efficient automated algorithms for the segmentation of medical images [16]. Some of the well-known Deep Neural Network (DNN) models include Convolutional Neural Networks (CNNs) [70], Recurrent Neural Networks (RNNs) [42] and Long-Short Term Memory (LSTM) [56].

Manual and semi-automated segmentation algorithms often rely on handcrafted features of the target anatomical structures in shallow learning [45]. Extracting such features is a tedious and expensive process that lacks a guarantee of optimality. Consequently, the selected features may not capture the comprehensive details of complex medical structures. Automated learning of relevant image characteristics by DL eliminates the manual feature engineering process and leads to robust feature representation [47]. Multiple layers of non-linearity enable DL to capture variability among different patients and conditions. DL based automated approaches are seen to have equivalent performance and even surpass humans in producing high-quality segmentation maps. They efficiently handle the huge volume of medical data regularly generated for timely diagnosis, thereby, improving the throughput of healthcare systems [87]. Thus, the applicability of DL in medical image segmentation has become an active area of research [111], [94].

The design of CNNs takes advantage of the local spatial connectivity of the neighboring pixels in an image [23]. Learning the strong correlation between adjacent pixels helps the network capture relevant features of the target region, such as edges, corners, and texture. A significant advancement in developing efficient algorithms to segment medical images was the development of Fully Convolutional Networks (FCN) [119] and the $U-$Net [106]. In addition to hierarchically modeling the complex patterns associated with anatomical structures with its encoder-decoder framework, $U-$ Net focused on pixel-level localization of target structures for efficient segmentation. The high-resolution spatial details of the target structure were combined with the overall contextual information modeled by the decoder via skip connections. This efficiently localized the position of the target anatomy in the entire image volume. The $U-$Net exhibited impressive results even with small datasets, to circumvent the common problem faced in the medical domain. The success of $U-$Net made it a benchmark in the field of medical image segmentation. Skip connections and

symmetric encoder-decoder became pivotal to the development of subsequent state-of-the-art (SOTA) frameworks, *viz.* $V-$Net [85], Attention $U-$Net [97], $U-$Net++ [140], $U-$Net 3+ [60].

While the convolutional framework of U-Net helped capture spatial details and multi-scalar information, the fixed-size kernels with localized view made handling of long-range dependencies difficult. Modeling global contextual information is necessary to interpret the complex structure of different organs or tissues for their robust segmentation. Thus, Vision Transformers (ViT) [34] were integrated with CNN to capture detailed spatial relationships along with global contextual information. The combination of Transformers with CNN led to the development of a new class of hybrid SOTA architectures, *viz.* TransUNet [19], UNETR [53], Swin UNETR [52], TransAttUNet [18], DS-TransUNet [73]. Transformers use the self-attention mechanism [129] to capture global dependencies throughout the input image slice/volume. However, the quadratic computational complexity of self-attention incurs a huge computational requirement to process volumetric medical images. This motivated the exploration of alternatives to Transformers with linear computational complexity, *viz.* State-Space-Models (SSMs) such as Mamba [50] and Vision-xLSTMs [3].

The current SOTA approaches, although effective, suffer from the curse of high computational demands. It makes their deployment challenging in resource-constrained settings. Healthcare facilities in developing or underdeveloped countries lack an advanced computing infrastructure to execute advanced algorithms while diagnosing medical images [4]. This necessitates the development of resource-efficient algorithms for the wide adaptability of DL frameworks to diverse healthcare settings; ranging from large centralized hospitals to point-of-care environments. The purpose of the thesis is to address this crucial requirement by designing and implementing a series of novel DL-based frameworks to accurately segment medical images in a resource-efficient manner. The algorithms developed are evaluated on multiple medical imaging modalities (both 2D and 3D). The research systematically explores the refinement of CNNs with advanced attention mechanisms to develop efficient hybrid architectures for medical image segmentation.

1.2 Motivation

Despite the noticeable accuracy achieved by the DL methods in segmenting medical images [96], their actual deployment in real-world settings is often challenging. The primary motivation for this thesis emerges from the significant computational demands and energy consumption demanded during the execution of the algorithms for inference on high-dimensional medical images.

Traversing deeper into a CNN architecture increases the effective receptive field [121]. This allows the model to acquire relevant features from a larger area. Additionally, a larger number of convolutional layers learn complex abstract features of target structures hierarchically from low-level image features [123]. These make deeper CNNs favorable for image-related tasks, such as medical image segmentation. However, it results in a large parameter-heavy model.

Storing and retrieving such parameter-intensive models for inference, especially for large volumetric images, poses a huge memory requirement.

ViTs, on the other hand, have a comparatively lower number of parameters than CNNs. However, they entail a substantial number of Floating Point Operations (FLOPs) [43]. This is due to the application of the self-attention mechanism on high-dimensional images. Firstly, the images are divided into multiple patches, which are later flattened and transformed into tokens. The attention score of each token is calculated using matrix multiplications, to represent the strength of its relation to all other tokens in the sequence. The number of patches or tokens (N) increases with a higher dimension of the input image. This increases the required number of pairwise comparisons while computing the attention scores; thereby, quadratically scaling the self-attention mechanism with respect to the number of tokens [$\mathcal{O}(N^2)$]. Consequently, powerful and expensive Graphical Processing Units (GPUs) and Tensor Processing Units (TPUs) become necessary to train and infer from ViT-based models.

The high computation involved in training and deploying these DL models has a significant impact on the environment [31]. The computational power needed to train parameter-heavy models leads to high electricity consumption. Although a single inference operation requires a lower amount of energy than training, the cumulative amount of energy needed to process the increasing number of high-dimensional medical images becomes significantly large. As electricity generation often depends on fossil fuels, it leads to enormous carbon emission. The powerful GPUs and TPUs generate immense heat during training and inference. Excess heat leads to degradation in hardware quality and shortens their lifespan. Therefore, an effective cooling system is necessary to ensure their optimal functioning. Large data centers equip water-based cooling systems to reduce the heat dissipated from processors. This type of cooling onsite results in a huge amount of water consumption that can lead to significant water withdrawal.

In addition to environmental concerns, a major challenge in integrating advanced DL models into the clinical workflow is the unavailability of high-end resources in different medical settings. Healthcare facilities in developing countries such as India face a budget and infrastructural crisis [27]. They lack the financial support to set up powerful computing clusters. Consequently, developing DL solutions for Point-of-Care (PoC) applications can be a suitable alternative to ensure wide accessibility of modern solutions at low cost. PoC healthcare facilities can provide rapid diagnostic results in remote areas or for patients requiring bedside care [128]. This helps make urgent decisions in critical cases and eliminates the need for experts to physically visit diagnostic centers. PoC tools do not depend on sophisticated computing clusters and can analyze captured images on local hardware. Therefore, computationally efficient DL algorithms are needed for smooth execution on PoC tools, which are usually equipped with less processing power and memory.

The computational demands of current algorithms, their environmental impact, and limited computing power drive the exploration of novel, resource-efficient, and accurate segmentation strategies. This thesis, in its search for efficient DL-based medical image segmentation algorithms, addresses the gaps

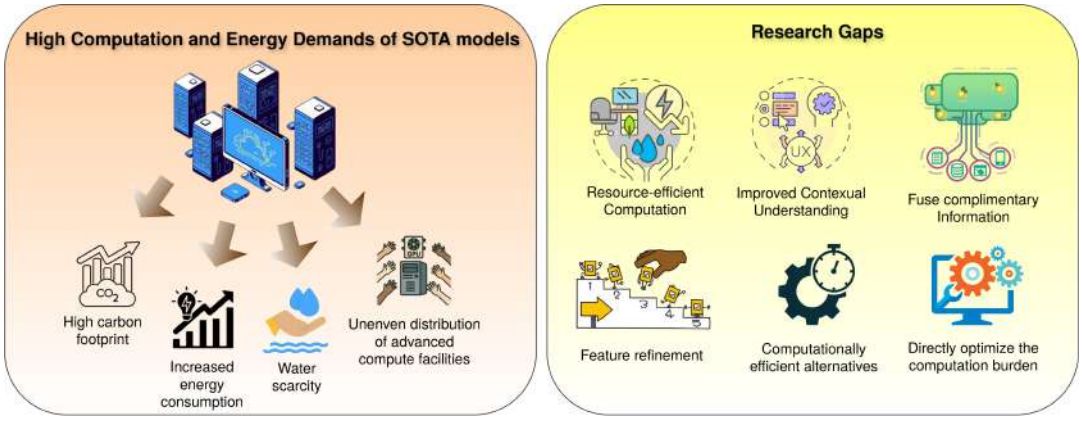

FIGURE 1.1: Motivation for developing efficient DL models for medical image segmentation.

as illustrated in Fig. 1.1.

- Need for resource-efficient DL architectures for medical image segmentation, with enhanced feature representations to handle both 2D and 3D image modalities.
- Requirement for integrating complementary information from spatial and spectral domains, with lower computational overhead; thereby, leading to richer feature representations.
- Computationally efficient alternatives to ViTs for efficiently modelling dependencies across multiple dimensions in volumetric data.
- Exploring techniques to reduce the computational burden associated with ViTs.

1.3 Preliminaries

This section outlines the fundamental concepts of image analysis and DL, to lay the foundation of the technical contributions discussed in the following chapters. A concise overview of images, their segmentation, and the role of spectral domain information are described. Then it elaborates several fundamental elements of DL, including MLP, CNN, and the $U-$Net framework, along with variants. Finally, an overview of some recent advances, *viz.* Vision Transformers and Vision-xLSTMs, is provided.

1.3.1 Image analysis

A digital image is mathematically expressed as a 2D matrix $I \in \mathbf{R}^{m \times n}$, where each element $I^{(i,j)}_{0 \leq i < m, 0 \leq j < n}$ denotes the intensity of the pixel in the location (i, j) [46]. The pixel intensities are in the range $[0, 255]$, representing gray shades ranging from black (0) to white (255) in grayscale images.

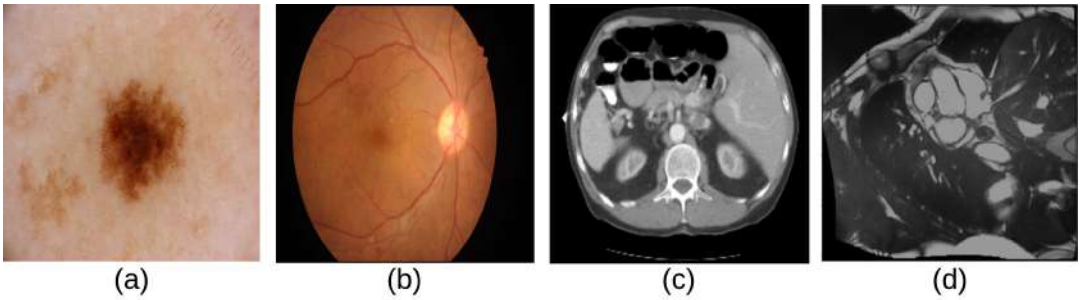

FIGURE 1.2: Sample medical images from different modalities. (a) Dermoscopy of skin lesion, (b) color fundus of retina, both in RGB color model; (c) axial CT slice of abdomen, and (d) cardiac MRI slice.

Color images are composed of three channels, corresponding to each of the primary colors, Red, Green, and Blue (in the RGB color model). The intensity of an element in each of these channels lies in the range $[0, 255]$. The value of the pixel at the location (i, j) is determined by a combination of the values of each of the three channels at that location. Sample medical images, including dermoscopic, color fundus of the retina, 2D abdominal CT slice, and cardiac MRI are illustrated in Fig. 1.2. The color fundus and dermoscopic images follow the RGB color model. The intensities of the voxel (pixel in 3D) in volumetric medical images, such as CT and MRI, are in the range [-1000 to +3000] to distinguish between bones and different tissues.

Segmentation: It includes a collection of techniques that divide the spatial region of an image I into n sub-regions defined by the set $S = \{I_1, I_2, \ldots, I_n\}$. The elements in S must satisfy the conditions [46]

1. $\forall I_i \in S, (Connect(I_i) = TRUE)$,
2. $\forall I_i \in S, (\bigcup_{i=1}^{n} I_i = I)$,
3. $\forall I_i, I_j \in S, (i \neq j \implies I_i \bigcap I_j = \phi)$,
4. $\forall I_i \in S, (P(I_i) = TRUE)$,
5. $\forall I_i, I_j \in S, (i \neq j \wedge Adj(I_i, I_j) \implies P(I_i \bigcup I_j) = FALSE)$,

where $Connect(I_i), P(I_i), Adj(I_i, I_j)$ are the predicates defined in I_k and ϕ is the null set. Here $Connect(I_i)$ is $TRUE$ if the sub-region I_i is a connected component, and $Adj(I_i, I_j)$ is $TRUE$ when I_i and I_j are adjacent sub-regions.

Condition 1 asserts that all the pixels in any sub-region must be connected (for e.g. by 4-connectivity or 8-connectivity). According to condition 2, the union of all the sub-regions must be equivalent to the image I. This concludes that every pixel in I must belong to a sub-region. Besides, a pixel cannot be a member of more than one sub-region, as defined by condition 3. This ensures exclusivity regarding the membership of a pixel to a sub-region. $P(.)$ signifies the homogeneity of sub-region I_i, which means $P(I_i) = TRUE$ if the pixels $\in I_i$ follow uniformity in terms of properties such as intensity or texture (condition

4). Condition 5 implies that two adjacent regions are heterogeneous as they must vary with respect to the image properties.

Classical methods of image segmentation explicitly define the predicate $P(.)$ based on hand-crafted features, whereas in DL the $P(.)$ becomes an implicit function. DL methods assign every pixel a single class label, thus satisfying the completeness (condition 2) and disjointness (condition 3) constraints. For any region I_i, $P(I_i) = TRUE$ means that all pixels in I_i have been assigned the same semantic class based on some common complex patterns between them (condition 4). The network learns an effective decision boundary to differentiate between adjacent regions that belong to different semantic classes (condition 5).

Spectral domain: Here, an image is represented in terms of its constituent spatial frequencies. It describes the rate of change of pixel intensities across the image from slowly varying regions (areas with uniform texture) to abrupt changes corresponding to edges and other fine details [93]. Discrete Wavelet Transform (DWT) is a form of spectral representation that decomposes the image into different frequency sub-bands using wavelets.

Wavelets are a family of waves, derived from a basic short wave (called the mother wavelet $\psi(t)$), to analyze images [46]. The mother wavelet is compressed or stretched along the time axis to generate its different variants, denoted by $\psi_{s,\tau}(t)$. These versions are slid along the image to analyze it piece by piece. Mathematically, wavelets are defined as

$$\psi_{s,\tau}(t) = 2^{s/2}\psi(2^s t - \tau). \tag{1.1}$$

Here, the parameter s is the integer scaling factor that controls the expansion or shrinkage of the mother wavelet. As the value of s increases, $2^s t$ becomes a more significant term; this results in the wavelet variant $\psi(2^s t - \tau)$ being a narrower version of $\psi(t)$. They are suitable for analyzing high-frequency features of an image, such as sharp edges, fine textures, and intricate details. On the other hand, a smaller value of s generates a wider wavelet variant useful for analyzing low-frequency features, *viz.* continuous smooth areas with uniform texture. The term $2^{s/2}$ is responsible for the amplitude of the resulting wavelet. The τ governs the amount of shift of the wavelet variant along the time axis. Changing τ, the wavelet variant of $\psi(t)$ is slid across different parts of the image. Each element of this set of wavelets, derived from $\psi(t)$, is fine-tuned to process a specific type of feature present at a particular location of an image.

DWT decomposes an input image I into a set of coefficients. This is carried out by expressing I as a linear combination of basis functions obtained from the scaling function $\alpha(t)$ and derived wavelets $\psi_{s,\tau}(t)$. Like wavelets, a family of scaling functions is generated from $\alpha(t)$ through scaling (controlled by the scaling parameter s) and shifting (controlled by the shifting factor τ). These capture smoother and low-frequency components at different scales (resolutions) and locations of I. The coefficients associated with the scaled and shifted scaling functions are referred to as approximation coefficients. The derived wavelet functions capture the high-frequency features. The coefficients associated with the derived wavelet functions are called detailed coefficients. Therefore, the

DWT decomposes I into different detailed and approximation coefficients on different scales.

1.3.2 Deep learning

Progress in neural learning significantly contributed to the automated analysis of images. This paradigm is conceptually derived from the cognitive processes of the human brain in relation to information processing. ANNs automatically extract the underlying patterns in a given dataset by adaptively modifying the strength of connections between simple processing units. The perceptron, a foundational element of neural learning, demonstrated remarkable success in its ability to learn decision boundaries to classify data samples [108]. However, the inability to model the non-linear decision boundary in the XOR problem established its limitations [86]. This served as a major stumbling block for research in neural learning.

MLPs were introduced [110] as a solution to this problem. An MLP develops a network of neurons involving a non-linear activation function, connected in layers, as illustrated in Fig. 1.3. The architecture models complex non-linear decision boundaries, enabling it to approximate any continuous function [57].

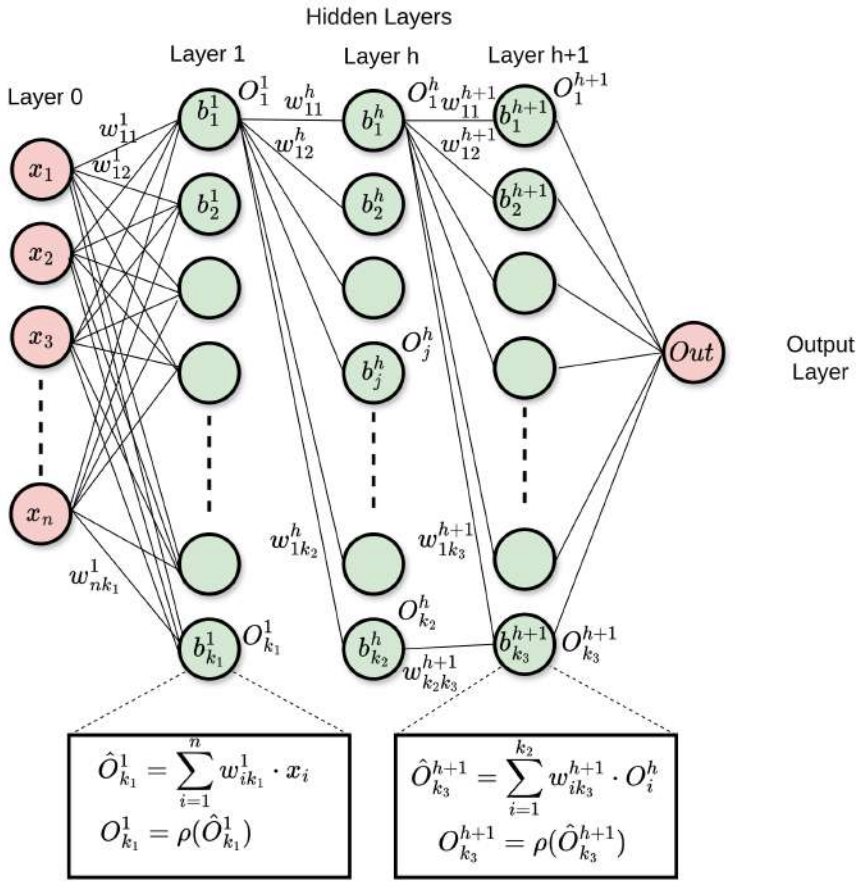

FIGURE 1.3: Schematic representation of an MLP.

The output of neuron j in layer h is the pre-activation value. It is a linear transformation of inputs from layer h. A simple stacking of multiple layers

of neurons will still result in a linear transformation of the inputs, regardless of the number of layers in the stack. The activation function ρ introduces non-linearity in neural output, to help it model complex, non-linear decision boundaries. This is propagated to neurons in subsequent layers in the forward pass. The final output Out of the network is thus a non-linear transformation of the input features X. Some examples of popular activation functions include Sigmoid [110], ReLU [95], Leaky ReLU [79], ELU [24], and GELU [55]. Let $X \in \mathbf{R}^{m \times n}$ denote the collection of m samples, each having n features, with $W^h \in \mathbf{R}^{k_1 \times k_2}$ and $B^h \in \mathbf{R}^{1 \times k_1}$ being the weight and bias matrices, respectively, at layer h. The forward pass of an MLP is given by

$$Out = W^{h+1}[\rho\{W^h\{\rho(W^2.\rho(W^1.X + B^1) + B^2)\} + B^h\} + B^{h+1}. \tag{1.2}$$

Initially, all weights and biases are assigned small random values. The objective is to find the optimal values of the weights and biases to minimize error during the prediction. The backpropagation of errors [110] achieves this goal by iteratively adjusting the weight and bias values based on the prediction error E. The essence of backpropagation is to compute the gradient of the error function with respect to each weight and bias in the network. Gradient computation requires the backward propagation of error signals from the subsequent layers, with the weight update governed by

$$w^h_{ij_{new}} = w^h_{ij_{old}} - \eta \frac{\partial E}{\partial w^h_{ij_{old}}} = w^h_{ij_{old}} - \eta (\frac{\partial E}{\partial \hat{O}^h_j} \times \frac{\partial \hat{O}^h_j}{\partial w^h_{ij_{old}}}), \tag{1.3}$$

where w^h_{ij} is the weight associated with the link connecting the neuron i of the previous layer to the neuron j of layer h, $\frac{\partial E}{\partial \hat{O}^h_j}$ is the rate of change in E w.r.t. the pre-activation output. This is the error signal e^h_j for neuron j in layer h, calculated as

$$e^h_j = \frac{\partial E}{\partial O^h_j} \times \frac{\partial O^h_j}{\partial \hat{O}^h_j}, \tag{1.4}$$

with $\frac{\partial O^h_j}{\partial \hat{O}^h_j}$ indicating the rate of change of post-activation output O^h_j w.r.t. $\hat{O}^h_j$, $\frac{\partial E}{\partial O^h_j}$ being obtained by aggregating the error signals of all neurons in the layer $h+1$, weighted by their connection strengths w^{h+1}_{jk}. This becomes

$$\frac{\partial E}{\partial O^h_j} = \sum_k e^{(h+1)}_k \cdot w^{h+1}_{jk}. \tag{1.5}$$

The error signal for the neuron j in layer h is

$$e^h_j = \left(\frac{\partial O^h_j}{\partial \hat{O}^h_j}\right) \left(\sum_k e^{(h+1)}_k \cdot w^{h+1}_{jk}\right). \tag{1.6}$$

Here $\frac{\partial \hat{O}^h_j}{\partial w^h_{ij_{old}}}$ represents the rate of change of the pre-activation output w.r.t. weight $w^h_{ij_{old}}$, η is the learning rate that controls the step size of gradient descent

to reach the minima of the error function. The algorithm terminates when the magnitude of consecutive update of the model weights becomes negligibly small.

However, with the availability of larger volumes of multimodal data over the Internet, the computational complexity of neural processing became infeasible. The design and development of ImageNet database [29] and GPUs changed the learning scenario in the big data framework, with the pioneering deep network AlexNet [67] successfully leveraging this combination to achieve remarkable human-like performance with image classification. This illustrated the efficacy of very deep neural networks, when combined with sufficient data and computational resources. It facilitated the DL revolution that later influenced various domains, including medical image analysis.

Convolutional Neural Network (CNN): Typically, it learns features from data arranged in a grid-like topology. Medical images are 2D/3D matrices with pixels/voxels arranged in rows and columns. Their effectiveness in processing image data makes them widely popular in computer vision applications. A generic CNN [70] consists of different types of layers *viz.* convolution, pooling, and a fully connected layer (MLP), as illustrated in Fig. 1.4 and is described below.

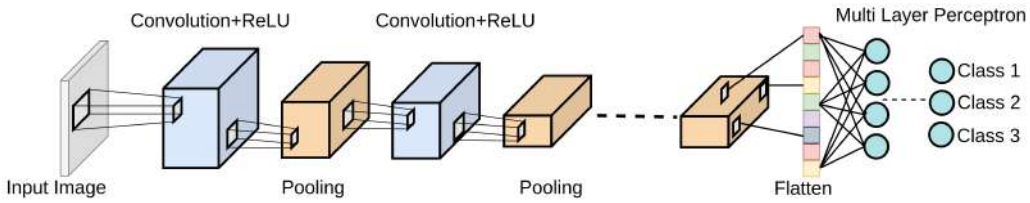

FIGURE 1.4: Schematic representation of a CNN.

- *Convolution layers:* These form the core of CNNs responsible for the extraction of features from an input image I. A bank of learnable filters/kernels is applied to the image. Each filter $f(.) \in \mathbf{R}^{m \times n}$ is a matrix of weights that traverses the entire image to extract relevant features (convolution operation). The pixel values within the region of the filter are multiplied element-wise by the filter weights. The resultant values are aggregated to produce a single output in the feature map. Mathematically, the convolution operation $(*)$ between the input image $I \in \mathbf{R}^{M \times N}$ and $f(.)$ is expressed as

$$F(i,j) = (I * f)(i,j) = \sum_{x=0}^{m-1} \sum_{y=0}^{n-1} I(i+x, j+y) \cdot f(x,y). \tag{1.7}$$

Here, the indices i and j denote the top left corner of the image region overlapped by the filter, and $F(i,j)$ is the resulting output value in the feature map as a result of the dot product between $f(.)$ and the sub-region of I overlapped by $f(.)$. Popular variants of the convolution layer include

- *Dilated convolution:* They expand the receptive field of the convolution filter without increasing the total number of trainable parameters. Although a larger kernel size enables the filters to extract features from a wider sub-region of I, it leads to an increase in the number of arithmetical operations and inflates the computational complexity. In this context, the dilated convolution introduces gaps between the filter weights to expand the field of view without adding extra parameters. The gap size is controlled by the parameter r called the dilation rate. Mathematically, the dilated convolution operation is expressed as

$$F(i,j) = (I * f)(i,j) = \sum_{x=0}^{m-1} \sum_{y=0}^{n-1} I(i + rx, j + ry) \cdot f(x,y). \quad (1.8)$$

- *Depthwise separable convolution:* They are an efficient modification to the standard convolution operation, found suitable for resource-constrained environments. A standard convolution filter processes an entire sub-region I having dimension $m \times n$ along with C channels. The depthwise-separable convolution applies a separate filter to every channel independently, instead of combining information across different channels. The intermediate output obtained is processed by a 1×1 convolution filter to combine information across channels. This approach of factorizing the standard convolution operation leads to a significant reduction in the computational cost, without degrading the performance of the neural network.
- *Transposed convolution:* The output feature map of a standard convolution has a reduced spatial dimension or is the same (by using padding with the input) compared to the input. Transposed convolution does the opposite, by expanding the dimension of the input feature map. It is used for upsampling the intermediate outputs specifically for image segmentation, where the resolution of the final output must match that of the input image. In comparison to standard upsampling operations, like bilinear upsampling (which has predefined parameters), the convolution filters learn the optimal set of parameters to upsample the input. A single spatial location of the input is multiplied by the kernel weights to generate multiple output values on the upsampled feature map.

- *Pooling layers:* They help reduce the spatial dimensions of the input by retaining relevant information. Such reduction in dimensionality helps to reduce the total computations involved in the subsequent layers of the network, to make it fast and efficient. They make the learned representation immune to subtle translations or distortions in the input. For example, if a feature (such as an edge) has a minor displacement in the input, a pooling layer can still output a similar response; thereby, enhancing the robustness of the network. The pooling operation is carried out independently, in each channel of the input feature volume, by sliding a $m \times n$

dimensional window across the input. The different types of pooling operations typically used in CNNs are summarized here.

- Max pooling: At every window position, the maximum value is chosen as the corresponding output for that region. This is effective in retaining the most prominent feature (such as a bright edge) within that region.
- Average pooling: At every window position, the average of all values within that window is computed. This forms the corresponding output for that region. It is effective in computing a general representation of features within that region instead of focusing only on the dominant one.
- Global pooling: Instead of sliding a fixed-sized window across the entire input, this applies the max or average operation on the entire spatial region. The operation represents an overall summary of the entire feature map by a single output value.

CNNs with multiple stacked convolution layers promote hierarchical learning of complex shapes and structures from basic image characteristics, such as edges, while preserving the spatial relationship between them. The weight-sharing property reduces the number of parameters needed to process the images in comparison to MLPs. Furthermore, translation invariance of the pooling layers helps to detect similar features even when their positions change in the input. These characteristics make CNNs a popular choice for handling image data.

$U-$**Net:** Capturing rich semantic details while preserving fine-grained spatial information remained a significant challenge in accurately segmenting anatomical regions from medical images. The architectural design of $U-$Net [106], based on CNN, addressed the dual challenge effectively; with considerable improvement in medical image segmentation. The $U-$Net is a fully convolutional symmetric encoder-decoder architecture, as shown in Fig. 1.5.

The encoder path consists of several stages, each featuring stacked 3×3 convolutional filters. The volume of the output feature map of each stage undergoes a max pooling operation before being transmitted to the subsequent stage. The encoder gradually reduces the spatial dimension while increasing the number of feature maps throughout its path. This facilitates hierarchical learning of global contextual information from the fundamental features captured in the initial layers of the encoder. The decoder gradually upsamples the feature maps, using transposed convolutions, to generate the final segmentation. The success of $U-$Net is due to skip connections, which concatenate feature maps from each encoder stage to the corresponding decoder level. This guarantees effective propagation and integration of high-resolution spatial details from preceding layers with global semantic information acquired at the deeper layers.

The edges of the anatomical structures in medical images range from simple and sharp (such as cortical bone) to complex, ambiguous, and low contrast (such as the interface between soft tissue organs such as the liver and spleen).

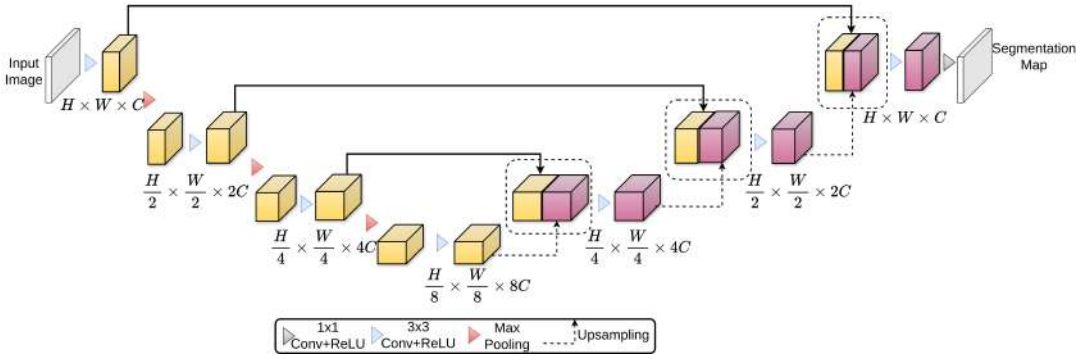

FIGURE 1.5: Schematic representation of $U-$Net.

Fine-grained spatial details play a crucial role in precisely detecting challenging boundaries. Skip connections prevent the minute details from getting lost in the deeper levels of the network and ensure that they become directly available to the decoder when localizing the target structures in the output. The combination of fine localized details with contextual information allows better reconstruction, of complex shapes and structures, at the output [106].

The feature maps generated in the final stage of the decoder undergo a 1×1 convolution to produce the final segmentation output. The total number of output channels corresponds to the total number of classes in the segmentation task. A softmax function is used on the convolution output in multi-class segmentation, while a sigmoid function is employed in binary-class segmentation to derive the class probabilities for each pixel or voxel.

The success of $U-$Net led to the development of several SOTA architectures that leveraged its core design principles. $V-$Net [85] is one of the early extensions of $U-$Net, designed specifically to handle volumetric images. It used the Dice loss function to address class imbalance, which remains a major issue in medical images due to the availability of a smaller number of pixels associated with target structures as compared to the background pixels. The $U-$Net++ [140] hypothesized that there exists a semantic gap between the encoder and decoder paths, because of which the direct fusion of feature maps might not be optimal. The model bridged this gap by redesigning the skip connections of $U-$Net. It implemented nested and dense skip connections with intermediate convolutional layers, enabling semantic enhancement of encoder features prior to fusion. The features of different stages of the encoder were efficiently aggregated in $U-$Net 3+ [60]. It aggregated feature maps from different levels of the encoder before integrating with the decoder feature maps at the corresponding level. This helped capture both fine-grained details and coarse-grained semantic information from feature maps across all levels of the encoder. These variants highlight the versatility of $U-$Net in performing automated medical image segmentation.

Another important aspect was strengthening the feature representation capacity of a model. A line of research aimed to make the networks focus more on relevant patterns while discarding the irrelevant ones. It was inspired by the concept of human visual attention, which filters relevant aspects from visual

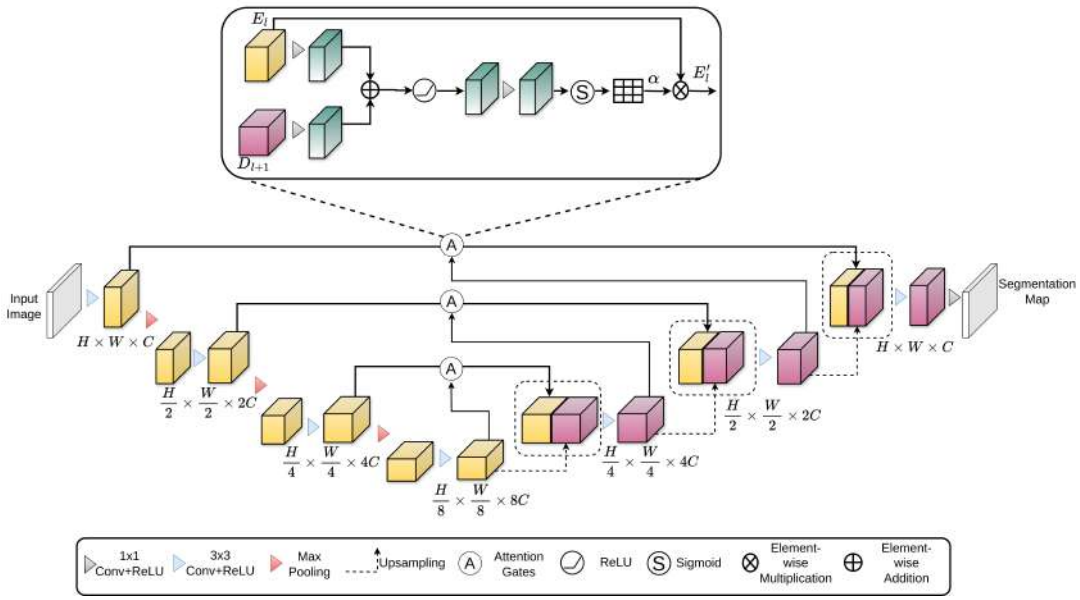

FIGURE 1.6: Schematic representation of Attention $U-$Net.

stimuli. This led to the incorporation of attention mechanisms into the segmentation models [88]. Such selective processing enabled efficient information processing and decision making. The network, instead of equally treating all the features, assigns higher weights to salient information while down-weighting the irrelevant ones. This is very helpful in medical image segmentation, where the target structures might be small, have variable shapes, or be surrounded by complicated or unimportant background details.

Attention $U-$Net [97] was developed by directly incorporating additional attention gates into the skip connections of vanilla $U-$Net, as shown in Fig. 1.6. The attention gates refine the feature maps of the encoder E_l before combining them with the decoder maps of the corresponding stage l. Feature maps from both the encoder (representing fine-grained details) and deeper levels of the decoder D_{l+1} (global abstract information) are provided as input to the attention gates. These gates initially subject both feature volumes to 1×1 convolution to equalize the number of feature channels. The resultant maps are added and passed through a ReLU activation, which emphasizes the prominent positive activations while dampening the smaller ones. Finally, the intermediate output is subjected to a 1×1 convolution followed by sigmoid activation to generate the weight map $\alpha \in [0, 1]$. The weights refine those encoder features that selectively emphasize relevant activations while attenuating the irrelevant feature responses. Consequently, the skip connections transfer relevant information to the decoder for fusion; thereby, leading to precise localization of target structures. The mechanism helps improve the sensitivity for smaller, less distinguishable structures by highlighting their features to improve overall segmentation performance.

Vision Transformer: The core component of Attention U-Net was the convolution operation, which inherently had a localized field of view. Although stacking multiple layers of convolution increased the effective receptive field,

explicit capture of long-range global dependencies remained a significant challenge for convolutional frameworks. For example, a small region of a liver might have a similar appearance to a small patch of a spleen in localized view. Capturing global dependencies can help the model distinguish between these different anatomical regions by understanding the overall context to correctly classify the pixels.

Computer vision researchers explored the applicability of a powerful neural network model, transformers, to address this core limitation [72]. Vision Transformers (ViT) [34], a modified variant of transformers for vision-related tasks, employed the self-attention mechanism to capture global contextual information. The input image was divided into fixed-size, non-overlapping patches. The self-attention mechanism parallelly modeled the relationships between every pair of patches, regardless of their spatial distance. This enabled ViTs to directly capture the global context of an input image. The key components of the model are described below.

- Generate patch embedding: Transformers are designed to process a 1D sequence of patch embeddings, called tokens. Since images are 2D matrices of pixels, they need to be converted into multiple sequences of 1D vectors. At first, the image of dimension $H \times W$ is divided into a grid of multiple small, non-overlapping patches. Each patch has a dimension of $P \times P$. Consequently, the 2D patches are flattened to a 1D vector. These flattened vectors are projected to a smaller embedding space by a learnable embedding matrix $E \in \mathbf{R}^{P^2 \times D}$. Patch embeddings, or tokens, form an abstract semantic representation of the patch contents. Furthermore, 1D-learnable positional embeddings ($E_{pos} \in \mathbf{R}^{\frac{HW}{P^2} \times D}$) are added to the tokens as

$$p' = [p^1.E, p^2.E, \dots p^{\frac{HW}{P^2}}.E] + E_{pos}, \tag{1.9}$$

  where $[p^1, \dots, p^{\frac{HW}{P^2}}]$ represents the collection of flattened patches, and $p' \in \mathbf{R}^{\frac{HW}{P^2} \times D}$ is the transformed set of tokens in the embedding space E.

- Self-attention mechanism: The patch embeddings in isolation are effectively static with no knowledge of their relationship with the surrounding regions in the image. The self-attention modules take these static patch embeddings as input and transform them into dynamic contextual embeddings. Every patch in the self-attention mechanism communicates with every other patch to build the global context.

  Every patch embedding is transformed to three vectors, *viz.* query ($Q \in \mathbf{R}^{\frac{HW}{P^2} \times D'}$), key ($K \in \mathbf{R}^{\frac{HW}{P^2} \times D'}$) and value ($V \in \mathbf{R}^{\frac{HW}{P^2} \times D'}$) by projecting them to three different subspaces using $W^Q \in \mathbf{R}^{D \times D'}, W^K \in \mathbf{R}^{D \times D'}$ and $W^V \in \mathbf{R}^{D \times D'}$ matrices. The similarity score to measure the relevance between different patches becomes

$$A = Softmax(\frac{Q.K^T}{\sqrt{D'}}). \tag{1.10}$$

Here, $A \in \mathbf{R}^{\frac{HW}{P^2} \times \frac{HW}{P^2}}$ is the similarity score matrix, with an element A_{ij} representing the amount of relevance between the ith and jth patches. A high similarity score value indicates higher relevance. The $Softmax(.)$ squashes scores in the range $[0, 1]$ to ensure that they sum up to 1.

The embedding of all the patches is now updated to incorporate the contextual information from all the related patches in the input image as

$$p'' = A.V, \tag{1.11}$$

with p'' being the updated patch embedding to encapsulate the global contextual information.

The above process is performed in parallel, multiple times, through multi-head attention. Each head attempts to model the global context from different perspectives. Each head contains a distinct set of W^Q, W^K and W^V matrices, which are updated during training.

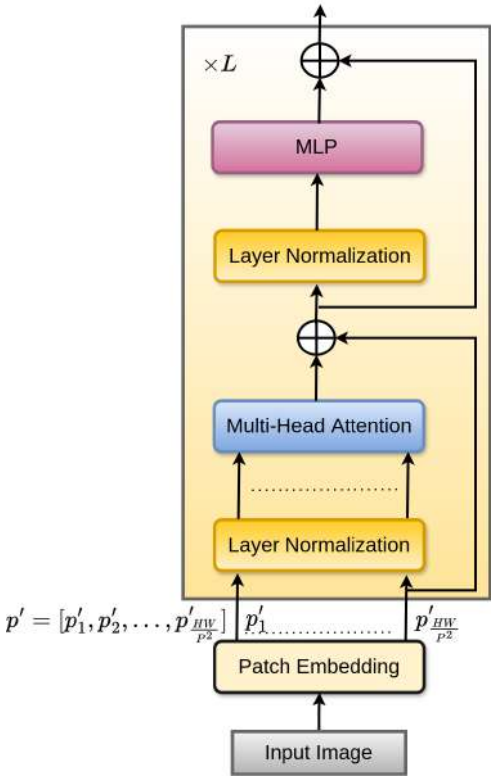

FIGURE 1.7: Schematic representation of an encoder block.

These key components of the ViT are assembled together within a single block of the encoder layer, as shown in Fig. 1.7. Here, L blocks are vertically stacked to progressively develop richer contextual representations. The patch embedding layer generates positional information-infused embeddings for each image patch. These embeddings are subsequently normalized by a layer normalization operation [8] prior to being processed by the multi-head attention layer. The generated contextual embeddings are fed to an MLP, which primarily serves to incorporate non-linearity into the patch representation. Skip

connections are added around the multi-head attention and MLP sub-layer, to mitigate the issue of vanishing gradients and ensure stable training.

The original ViT, initially designed for image classification, has been widely adapted for image segmentation due to its powerful feature extraction abilities. Contextual embeddings are transformed into spatial feature maps, which are fed to a decoder path for progressive upsampling for the final segmentation output.

Vision-xLSTM: The powerful global context modeling of Vision Transformers with self-attention has made them a popular choice for various vision-related tasks. However, quadratic memory and computational complexity ($O(N^2)$) with respect to the total number of image patches pose a significant challenge in their application to high-dimensional medical images. This limitation fuels the search for computationally efficient alternatives to ViTs.

The need for computational efficiency has revived interest in recurrent networks, such as Long Short Term Memory (LSTM) [56], which preceded transformers. Recently, LSTMs have been extended to a high-performance and scalable model called extended LSTM (xLSTM) [11]. xLSTMs overcome the limitations of LSTMs through key innovations *viz.* exponential gating and parallel matrix memory. They exhibit a higher potential with respect to computational and memory efficiency in comparison to ViT and State Space Models (SSM) [50]. Vision-xLSTM (ViL) [3] pioneers the application of xLSTMs to computer vision tasks.

The input image is transformed into patch embeddings analogous to ViT. The flattened patches, after normalization, are projected onto an embedding space to increase their dimension by a factor of 2. These expanded embeddings are divided into two paths, $x_{mlstm} \in \mathbb{R}^{\frac{HW}{P^2} \times 2Z}$ and $y \in \mathbb{R}^{\frac{HW}{P^2} \times 2Z}$ as shown in Fig. 1.8. x_{mlstm} is further processed by the mLSTM layer of the ViL block.

The mLSTM [11] layer depicted in Fig. 1.8 is responsible for modeling the inter-patch dependencies. It is an enhanced variant of LSTMs that features a matrix memory cell state $\mathbf{C_t}$ rather than a scalar value. At a given time step t, x_{mlstm_t} undergoes a 1D causal convolution with SiLU activation [41]. The intermediate result ($\mathbf{X_t} \in \mathbb{R}^{N \times 2Z}$), which is the current input, is then mapped onto query (Q_t), key (K_t) and value (V_t) vectors.

The generation of query, value and keys is mathematically represented as

$$\mathbf{Q_t} = \mathbf{X_t}\mathbf{W}_{Q_t}^T, \quad \mathbf{K_t} = \mathbf{X_t}\mathbf{W}_{K_t}^T, \quad \mathbf{V_t} = \mathbf{X_t}\mathbf{W}_{V_t}^T, \tag{1.12}$$

where $\mathbf{Q_t} \in \mathbb{R}^{\frac{HW}{P^2} \times 2Z}$, $\mathbf{K_t} \in \mathbb{R}^{\frac{HW}{P^2} \times 2Z}$,$\mathbf{V_t} \in \mathbb{R}^{\frac{HW}{P^2} \times 2Z}$ are the query, key, and value matrices, respectively. Here, $W_{Q_t} \in \mathbb{R}^{2Z \times 2Z}$, $W_{K_t} \in \mathbb{R}^{2Z \times 2Z}$, $W_{V_t} \in \mathbb{R}^{2Z \times 2Z}$, are learnable weight matrices to generate the query, key, and value vectors.

The input and forget gate pre-activations, $i_t \in \mathbb{R}^{\frac{HW}{P^2} \times 2Z}$ and $f_t \in \mathbb{R}^{\frac{HW}{P^2} \times 2Z}$ are calculated simultaneously from $\mathbf{X_t}$ as

$$i_t = \exp((\mathbf{W^I})^T\mathbf{X_t} + \mathbf{B}), \tag{1.13}$$

$$f_t = \exp((\mathbf{W^F})^T\mathbf{X_t} + \mathbf{B}). \tag{1.14}$$

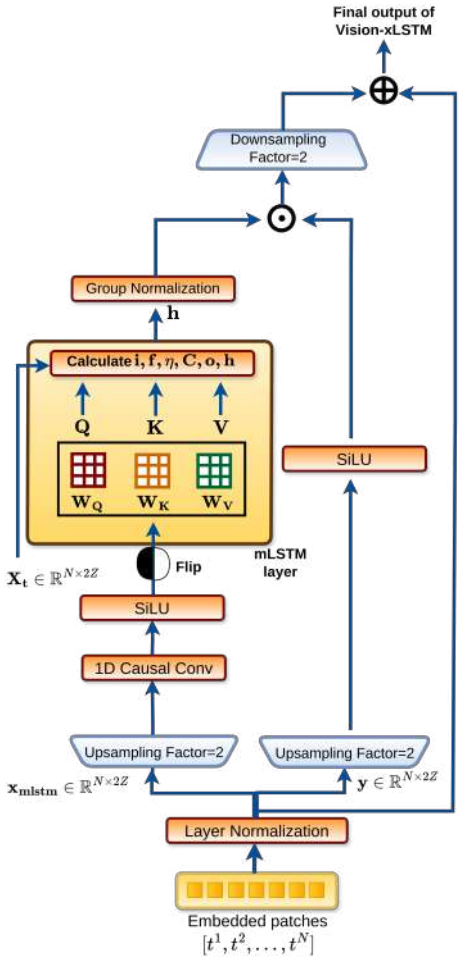

FIGURE 1.8: Architectural framework of mLSTM layer. The ViL module is made up by stacking L number of mLSTM layers to compute inter-patch dependencies.

Here, $\mathbf{B} \in \mathbb{R}^{\frac{HW}{P^2} \times 2Z}$ is the bias matrix. $\mathbf{W^I}, \mathbf{W^F} \in \mathbb{R}^{2Z \times 2Z}$ are the projection matrices for i_t and f_t respectively. exp represents the exponentiation. The input gate decides which information to store in the current cell state t. The forget gate is responsible for discarding information from the previous cell state on the timestamp $t-1$. The cell state $\mathbf{C_t}$ is updated as follows:

$$\mathbf{C_t} = i_t.\mathbf{V_t}\mathbf{K_t}^T + f_t.\mathbf{C_{t-1}} \tag{1.15}$$

$\mathbf{V_t}\mathbf{K_t}^T$ is the key-value pair representing the current input information to be stored at time step t. An additional normalizer state n_t is introduced to deal with the exploding cell state values due to the exponentiation operation. The normalizer state pre-activation is computed as follows:

$$n_t = f_t.n_{t-1} + i_t.\mathbf{K_t} \tag{1.16}$$

The normalizer states, until time-step $t-1$, are scaled by f_t. This determines the influence of the previous normalizer states on the current normalizer state. The contribution of the input of the current state $\mathbf{K_t}$ is controlled by i_t.

Subsequently, the hidden state h_t is calculated as follows:

$$h_t = o_t[\mathbf{C_t Q_t}/\mathbf{max}\{|\mathbf{n_t^T Q_t}|, \mathbf{1}\}] \tag{1.17}$$

$\mathbf{C_t Q_t}$ corresponds to the retrieval of information from the current state of memory cell $\mathbf{C_t}$ using the query $\mathbf{Q_t}$. The information retrieved is normalized to n_t. $o_t = \sigma(\mathbf{W^{oT} X_t} + \mathbf{B})$ represents the activation of the output gate that selectively filters the information from $\mathbf{C_t}$ to be written to h_t. σ is the sigmoid activation function that squashes o_t in the range $[0, 1]$.

1.4 Literature Review

This section surveys the prominent DL architectures for medical image segmentation, building on the foundational concepts established in the preceding section. The review outlines the evolution of these models, starting with a summary of CNN models that integrate attention mechanisms to enhance feature representation. Subsequently, it analyzes the transition to ViT-based frameworks for an efficient capture of global context. Finally, it addresses the recent emergence of linear architectures, which have gained popularity due to their reduced computational complexity compared to ViTs. The key contributions and drawbacks of every approach are discussed, establishing a foundation for the novel and efficient solutions proposed in this thesis.

1.4.1 Incorporating attention modules in *U*-Net

The fundamental limitation of skip connections is that they transmit fewer discriminative features, which negatively impacts segmentation performance. Researchers incorporated additional attention modules into the *U*-Net to achieve precise segmentation of the target structure, with Attention *U*-Net being the benchmark architecture. These modules guide the model to focus on relevant spatial features for a precise segmentation of the target structures [26].

The spatial attention mechanism of Attention *U*-Net was expanded in [109] by introducing concurrent spatial and channel attention mechanisms. Channel attention identifies salient feature maps within the input encoder feature volume. The feature maps were squeezed into a vector by the global average pooling operation. Later, the vector was processed by an MLP to generate a channel attention map. D2A-Net [138] incorporated two novel attention modules, *viz.* the Gated Attention Module (GAM) and Decoder Attention Module (DAM) into the standard *U*-Net framework. GAM refined the encoder features along the skip connections using both channel and spatial attention mechanisms. It generated weights for recalibrating encoder feature maps, using feature maps from deeper layers (referred to as "guiding signals"). DAM was incorporated within the decoder to refine the upsampled decoder maps and suppress potential noise introduced from the upsampling method. The HADC-Net [21] introduced a dual hybrid attention strategy to tackle segmentation difficulties arising from significant variability in the shape and size of the target lesion. The encoder was

made up of the HADC encoder module, which simultaneously refined features across the convolution path and efficiently captured multi-scale attributes. The spatial attention map was made up of a dense and dilated convolution block to capture multi-scale information with an efficient flow of information. The squeeze and excite block [59] generated weight maps for channel recalibration.

Subsequent research enhanced the design of the attention module to capture multi-scale features. MiniSeg [102] introduced a lightweight attention module called Attentive Hierarchical Spatial Pyramid (AHSP) for the extraction of multi-scale features. AHSP comprised parallel branches with dilated depthwise separable convolutions having varying dilation rates. This facilitated efficient learning of both local characteristics and broader contextual understanding, with a reduced number of parameters. The feature maps from the concurrent branches were integrated and processed through a spatial attention mechanism. MA-UNet [15] placed the dual attention module at the segmentation head, which accepted input from multiple layers of the decoder. These feature maps captured relevant information at varying scales. The channel attention component of the dual attention module computed the impact of each channel on every other channel through matrix multiplication between feature maps. The concurrent spatial attention component computed the weighted sum of features across all the spatial locations of the feature volume. The weight map was then multiplied by the input feature volume for recalibrating important spatial locations. The context extractor module in CE-Net [51] captured rich global semantic information from the bottleneck of the *U*-Net. Initially, the encoder feature maps were subjected to multiple cascading dilated convolution layers with increasing dilation rates. Subsequently, fused intermediate representations from different dilated convolution were passed to a residual multi-kernel pooling block comprised of max-pooling layers of varying kernel sizes. This allowed the model to capture global contextual information from multi-sized kernels.

Several drawbacks related to the design choices existed despite the improvement in segmentation performance achieved by these methods. Dilated convolutions were incorporated into the architecture of attention modules to obtain multi-scale information. Convolution kernels with larger dilation rates could result in a gridding effect [58]. Additionally, the sparse convolution kernels potentially overlooked fine-grained details of the input. This affected the segmentation performance of small anatomical structures. The use of global average pooling on feature maps, to generate weights for channel recalibration, aggressively abstracted the overall spatial information. Equal importance was provided at each spatial location, and diminished the relevance of localized activations in the feature map. It occurred because of the averaging out of high activations in the smaller regions, to minimize their impact on the descriptor vector. Using multiple attention modules, similar to D2A-Net and CE-Net, often increased the computational burden and parameter count of the entire network. Complex multi-input fusion approaches created complicated gradient pathways, making the training process less stable.

1.4.2 Integrating ViT with CNN

The ability of CNNs to represent features is constrained to local spatial details because of their fixed-size receptive fields. Conversely, although ViTs are proficient in modeling global context, they are ineffective at capturing fine-grained details due to the loss of low-level spatial information during tokenization. Such complementary characteristics of CNNs and ViTs motivated the development of hybrid models to create a robust representation of anatomical structures [75]. TransUNet [19] pioneered leveraging the strengths of CNNs and ViTs for medical image segmentation. The CNN backbone in the encoder hierarchically constructed global abstract features from high-resolution low-level features. Multiple stacked blocks of ViTs at the bottleneck modeled global long-range dependencies on the feature maps from the encoder. A CNN upsampler decoder reconstructed the segmentation mask by employing skip connections to integrate the global context of ViTs with the detailed spatial information of the CNNs. TransAttUNet [18] enhanced this design by introducing the Self-Aware Attention module at the bottleneck to effectively model the spatial dependencies between two spatial locations, along with their global spatial relation. The self-attention component captured the global spatial context. Simultaneously, the Global Spatial Attention component quantified the impact of the pixel at the ith position on the pixel at the jth position.

Transfuse [137] introduced a novel parallel architecture, differing from the sequential bottleneck design. The input image was processed simultaneously with a ViT and CNN branch. A novel BiFusion module merged the output from both branches across multiple stages, using both channel and spatial attention mechanisms. Subsequently, hybrid models replaced the CNN-based encoder with a hierarchical Transformer backbone. UNETR [53], SegFormer [131], and Swin UNETR [52] illustrate this methodology, with SegFormer and UNETR employing multiple stacked ViTs and Swin UNETR utilizing Swin Transformers [76] as their respective encoders. Transformers directly learned features from the input, instead of using CNNs to hierarchically construct global abstract features from low-level spatial details. Consequently, global context was efficiently captured across different stages of the U-shaped segmentation model. DS-TransUNet [73] employed a dual-scale encoding scheme, with Swin Transformers along the encoder. Two parallel branches of Swin Transformers encoded the feature representation from the input with different patch sizes. This helped model both coarse-grained and fine-grained details, simultaneously, along both paths. The Transformer Interactive Fusion module fused the representation captured by both branches. UNETR++ [117] replaced the self-attention mechanism of UNETR with the Efficient-Paired Attention block to independently capture the spatial and channel dependencies. The spatial attention modeled global dependencies, with linear time complexity, by projecting keys and values to a lower-dimensional space. The channel attention modeled interdependencies between different channels.

Other popular approaches substituted both the encoder and decoder paths of the U-Net with Transformer blocks. SMESwin-UNet [130] employed Swin

Transformers in the encoder and decoder to hierarchically extract global abstract features, followed by gradual upsampling to generate segmentation output. Skip connections were replaced by a CNN branch that processed superpixels of the original input image to reduce interference from nearby structures. A novel Transformer-based attention block fused the features from the Transformer encoder with the CNN branch output. The CTC-Net [134] had a dual encoder structure, with one branch consisting of CNNs and the other made up of Swin Transformers. A fusion module integrated complementary features at multiple levels to model cross-domain correlation between the features. The decoder, built from Swin Transformer blocks, upsampled the encoder feature maps to generate the predicted output.

The family of Transformer-based segmentation models presents specific challenges. Models using the standard self-attention mechanism, namely TransUNet, UNETR, and TransAttUNet, exhibit quadratic computational complexity in relation to the dimensions of the input image. This results in a substantial increase in training parameters and computations. Multiple methods rely on pre-trained encoder backbones on natural image datasets. This leads to a suboptimal performance for medical image segmentation tasks, due to the domain gap between the natural images and medical image datasets. The addition of complex auxiliary modules, like fusion blocks for integrating CNN and Transformer features, adds to the computational burden. The tokenization process of image patches results in the loss of low-level spatial details. Consequently, it becomes challenging to segment small anatomical structures, especially using Transformer-only models.

1.4.3 Linear models for efficient segmentation

The significant computational demands of Transformers make it a challenging proposition to use for segmenting high-resolution medical images. A new research direction involves exploring alternative sequence models, like State Space Models (SSM) [50] or recurrent networks that have linear computational and memory complexity, to efficiently model global context. Initially developed for natural language processing tasks, SSM like Mamba was adapted for computer vision tasks to learn global dependencies with linear-time complexity. *U*-Mamba [78] incorporated hybrid CNN-SSM blocks at different stages of the encoder. Each CNN-SSM block extracted local spatial details with a CNN sub-module, followed by modeling of global dependencies in the feature maps using the Mamba sub-module. The decoder progressively upsampled the feature representation from the encoder with transposed convolutions to generate the segmentation output. SegMamba [132] introduced an improved version of CNN-SSM blocks called the Tri-Oriented Mamba (TOM) module for segmentation of volumetric medical data. TOM initially flattened the volumetric features along rows, starting from top-left, reversing the previous direction, starting from bottom-right, and then along the depth axis. The Mamba module was applied to each of the flattened sequences, followed by a combination of results for efficient capture of the volumetric relationship in the data. Unlike unidirectional

scanning of *U*-Mamba, the Swin UMamba [74] introduced a selective scanning method to unravel feature volumes along the four dimensions.

Other linear models, *viz.* Convolutional Long-Short Term Memory (ConvLSTM) [112] also enhance the performance of *U*-shaped segmentation frameworks. TBConv-LNet [62] introduced bidirectional ConvLSTM, with Swin Transformer blocks at the skip connections. The ConvLSTM detected temporal dependencies of the patterns in a bidirectional manner. Simultaneously, the Swin Transformer path captured the global spatial dependencies. The output of both modules was combined to generate a robust representation of the target structures. SEACU-Net [64] introduced a ConvLSTM based attention block at skip connections to refine the encoder feature maps, before combining with the decoder maps at the corresponding level. LC-UNet [65] followed a coarse-to-fine segmentation framework, with the encoder of fine segmentation network consisting of ConvLSTM blocks. The ConvLSTM-UNet [6] relied on ConvLSTM blocks to hierarchically model global abstract features, obtained from spatial details, to efficiently segment microscopy videos.

Linear models suffer from significant difficulties related to the precise segmentation of target structures while reducing computational demands. Mamba-based models fail to retain fine-grained details despite efficient computational and memory complexities while updating structured memory [84]. This results from the selective structured memory update mechanism that might lead to information loss. Consequently, it led to inefficient segmentation performance of small and localized anatomical structures. Additionally, Mamba faces difficulties in refining distant past information; which leads to inefficient long-range context correction. Processing of feature maps by flattening and scanning in a predefined pattern, as seen in *U*-Mamba or SegMamba, leads to loss of critical spatial information.

1.4.4 Redundancy reduction in ViT via token pruning

Researchers have explored model compression, like pruning irrelevant tokens from ViT, to tackle the substantial computational and memory demands. DynamicViT [103] incorporated lightweight modules to predict the relevance scores of tokens at each stage of the network. TRAM [81] modeled the self-attention mechanism through an MLP layer, with the tokens corresponding to the input nodes of the MLP. The association between different tokens was represented by the weights of the MLP. The relevance score of each token was computed recursively on the MLP representation of the self-attention mechanism. The tokens having the lowest scores were discarded at early layers to ensure that only relevant tokens were forwarded across the network. Evo-Vit [133] preserved the spatial structure of the token grid. It dynamically identified tokens, as informative or placeholder, based on a global class attention score. The placeholder tokens represented irrelevant ones, which were retained simply to maintain the grid structure of the attention map. They were aggregated into a single representative value, which was then processed by the self-attention mechanism to reduce computations.

Architectural modification of ViTs was also pursued to reduce the total number of tokens. Shunted Self-Attention [104] performed token reduction by merging across different attention heads to efficiently learn multi-scale representations. TA-ASF [20] implemented a two-stage token pruning method. An importance score was assigned to each token, followed by dividing the entire set of tokens into high- and low-importance sets. A subset of tokens was sampled from both groups, to preserve low-importance tokens that might be of global importance in the later layers of the network. Subsequently, the tokens in the subset were merged on their similarity. Soft-TopK token pruning [139] generated a score for each token with a lightweight module. Random noise was added to the scores to prevent the same tokens from being chosen repeatedly. The top-K score tokens were subsequently selected in the forward pass. The algorithm next applied a differentiable function to generate probabilities of selecting a token, which were then used to update the score prediction network through backpropagation.

Other investigators achieved efficiency by using alternate strategies. Intra Head Pruning [136] eliminated certain rows of the weight matrices, to produce query, key, and value vectors for the self-attention mechanism. An importance score was computed for each row to eliminate irrelevant entries. A supplementary relationship matrix was preserved to address dependency conflicts with subsequent layers. Token expansion [61] improved the training time of ViT by gradually growing the total number of tokens from a small set of initial seed tokens. Each stage introduced new tokens differing from the existing ones. The residual tokens were aggregated and processed along with the selected set. Early exit strategies [125], [127] save computation by bypassing the processing of specific layers for a subset of tokens.

The aforementioned pruning techniques suffer from drawbacks related to the lack of generalizability and computational overhead involved. Methods performing unstructured pruning hamper the grid-structure of the tokens; thereby, making them incompatible for hierarchical ViT models like Swin UNETR and UNETR. Hard token pruning strategies pose difficulty, due to the loss of essential fine-grained information required for reconstructing the segmentation output. Incorporating auxiliary modules, to calculate the token relevance scores, increases the complexity and parameter count of the model.

1.5 Performance Evaluation Metrics

This section provides an overview of the performance metrics used to evaluate the efficacy of the algorithms developed in this thesis.

Precision:
It measures the quality of positive predictions by quantifying the ratio of actual positive pixels (belonging to the target region) out of all the pixels labeled to be positive by the DL model. Mathematically, it is defined as

$$Precision = \frac{TP}{TP + FP}, \tag{1.18}$$

where TP and FP are the True Positives and False Positives, respectively.

Recall:
It defines how the model performs in correctly identifying all the pixels belonging to the target structure. It quantifies the ratio of predicted positive pixels out of all the pixels belonging to the target region in the ground truth. We have

$$Recall = \frac{TP}{TP + FN}, \tag{1.19}$$

where FN denotes the False Negatives.

Intersection-over-Union (IoU):
It measures the overlap between the predicted segmentation map and the ground truth by quantifying the ratio of the intersecting area between the predicted mask and the ground truth to the union area of both masks. We use

$$IoU = \frac{TP}{TP + FP + FN}, \tag{1.20}$$

with $IoU \in [0, 1]$. An IoU value of 0 represents no overlap and a value of 1 indicates a perfect overlap.

Dice Score Coefficient (DSC):
It measures the harmonic mean of $Precision$ and $Recall$ by generating a single value to define how the model has struck a balance in identifying all positive pixels ($Recall$) while making fewer errors ($Precision$). The DSC value belongs to the range $[0, 1]$, and is defined as

$$DSC = \frac{2 \times TP}{(2 \times TP) + FP + FN}. \tag{1.21}$$

95th percentile Hausdorff Distance (HD95):
It corresponds to the largest distance between the predicted region boundary ($\hat{\gamma}$) and the ground truth boundary (γ). The computation evaluates all the distances from every point on the predicted segmentation boundary to the closest point on the actual region boundary and returns the 95th percentile of those distances. This helps prevent sensitivity to outliers. Mathematically, we get

$$HD95_{\beta}(\hat{\gamma}, \gamma) = \max_{\beta \in c} \{ \max_{\hat{y}_{\beta, i \in \hat{\gamma}}} \min_{y_{\beta, j \in \gamma}} d(\hat{y}_{\beta, i \in \hat{\gamma}}, y_{\beta, j \in \gamma}), \max_{y_{\beta, i \in \gamma}} \min_{\hat{y}_{\beta, j \in \hat{\gamma}}} d(\hat{y}_{\beta, i \in \hat{\gamma}}, y_{\beta, j \in \gamma}) \}, \tag{1.22}$$

where c denotes the total number of segmentation classes and i corresponds to a pixel (or voxel), with $d(.)$ being the Euclidean distance.

1.6 Datasets and Software Packages

This section presents an overview of the various datasets, related to the medical image segmentation task, used to evaluate the performance of the algorithms

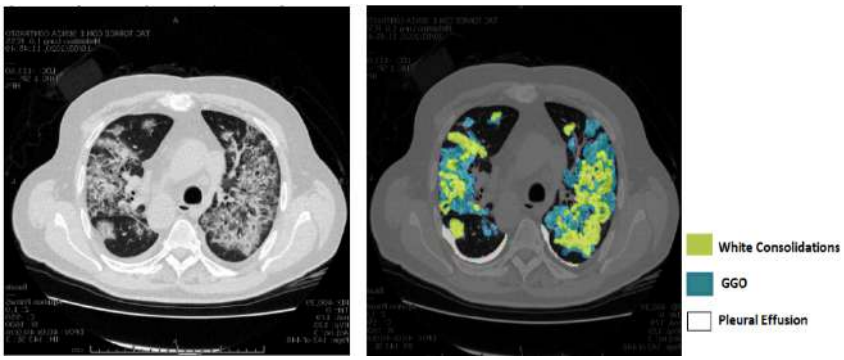

FIGURE 1.9: Sample lung CT slice with COVID-19 pathologies.

introduced in this thesis. We also present a summary of the various software packages employed to execute the proposed models. All models were trained on an NVIDIA RTX A6000 GPU with 48GB of RAM.

1.6.1 Datasets

The datasets used in different chapters of this thesis correspond to COVID-19 infection from lung CT scans, Diabetic Retinopathy from fundus images of the eye, Adrenal gland tumor from CT, Abdominal organs in CT, Skin lesions from dermoscopic images and Cardiac organs from chest magnetic resonance images.

COVID-19:
Lung CT slices with annotations for different COVID-19 pathologies were obtained from four publicly available datasets, as shown in Table 1.1. Conspicuous ground glass opacity (GGO) and multiple mottling lesions in the peripheral and posterior lung regions on CT images are hallmark characteristics of COVID-19 pneumonia [48], as shown in Fig. 1.9. GGOs are darkened hazy spots in the lung, diffuse enough not to block the underlying blood vessels or lung structures. The consolidations correspond to areas of increased lung density [48]. It is observed that with time these characteristics become more frequent and are likely to spread across both lungs.

Dataset-1 [92] constitutes lung CT scans taken from 1110 patients belonging to five categories *viz.* CT-0, CT-1, CT-2, CT-3, CT-4. Although CT-0 corresponded to normal cases, the rest of the data was divided into four groups according to the increased severity of infection in the lung. Here, CT-1 has 684 samples with an affected lung percentage of 25% or less, and CT-2 has 125 samples with affected areas ranging from 25% to 50%. CT-3 consists of 45 samples that have 50% to 75% of the affected lung region, and CT-4 has just two samples with 75% and above the affected lung portion. Only 50 scans, belonging to CT-1, were annotated with binary masks that represent regions of GGO. The affected lung area was assigned the label “1" and the rest of the slice (unaffected portion, without the lesion, along with the background region) was assigned the label “0". CT volumes with annotated masks were used for our study. Dataset-2 [82] consists of 100 axial CT slices from >40 COVID positive patients. The slices were labeled by a radiologist to demarcate three different

TABLE 1.1: Datasets used for segmenting COVID-19 lesions.

No.	Name	No. of annotated samples	No. of slices with lesions
1	MOSMED [92]	50	785
2	MedSeg-COV-1 [82]	>40	100
3	MedSeg-COV-2 [83]	9	373
4	COV-CT-Lung-Inf-Seg [77]	10	1351

pathologies of COVID-19 *viz.* GGOs, white consolidations, and pleural effusion. Dataset-3 [83] includes nine volumetric lung CT scans obtained from the Italian Society of Medical Interventional Radiology. However, of a total of 829 slices, only 373 slices were provided with annotations indicating the regions with GGOs and white consolidations. Dataset-4 [77] is a collection of lung CT scans from 20 patients with annotations performed by two radiologists. Later, these were validated by another experienced radiologist. The ground truth of these slices consisted of only two labels *viz.* "1" and "0", indicating the diseased tissues and other regions (comprising healthy regions of the lung and background). Here, we used lung CT volumes from the first ten patients to extract the slices in our experiments. This was because the remaining 10 samples contained a non-uniform number of slices, indicative of dissimilarity in the voxel spacing.

Diabetic Retinopathy (DR):
The IDRID dataset [101] has a collection of 81 color fundus images captured by a non-mydriatic camera with a 50-degree view. It was annotated for different DR pathologies – Microaneurysms, Hemorrhages, Hard Exudates, Soft Exudates, and Optic disc, as depicted in Fig. 1.10. Microaneurysms and hemorrhages, together known as Red Lesions, are primary indicators of DR. Microaneurysms are small balloon-shaped enlargements of the blood vessels of the retina. Hemorrhages are blood vessel ruptures that result in blood flow to the inner layer of the eye. Hard exudates are yellow spots, whereas soft exudates correspond to the white area with ill-defined edges on the retina. Here annotations for Microaneurysms and Hemorrhages together form the annotation for Red Lesions.

Each of the images in the IDRID dataset has a dimension of 4288×2848. The Messidor dataset [28] contains 1200 color fundus images captured by a non-mydriatic camera with a 45-degree view. Each image was cropped to a dimension of 512×512 that contained only the color fundus image. Of these 1200 images, 470 were properly annotated for the occurrence of Red Lesions, by our medical expert. The images of both datasets were individually, randomly classified into the train-test split by 80% and 20% of the original data.

Adrenal gland tumor:
Adrenocortical carcinoma (ACC) [91] is a rare tumor that originates in the adrenal cortex, which constitutes the external region of the adrenal glands. ACC is characterized by its high level of aggression and fatality, as is evident from its overall 5-year survival rates (varying between 14% and 44%). The dataset consists of contrast-enhanced CT imaging studies of 53 patients diagnosed with ACC, with the voxel-level segmentation of the tumors performed

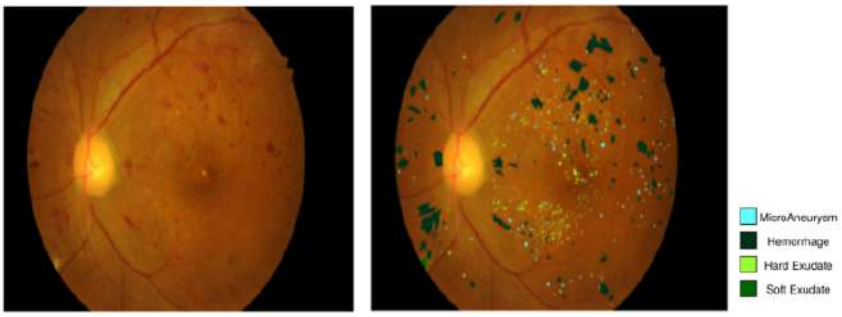

FIGURE 1.10: Sample fundus image illustrating DR pathologies.

by an experienced radiologist. The dimensions of the CT scan volumes ranged from $512 \times 512 \times 32$ to $512 \times 512 \times 140$.

Abdominal organs:
The dataset *Synapse* [68] consists of 30 CT volumes, with varying volume sizes ranging from $512 \times 512 \times 85$ to $512 \times 512 \times 198$. The CT volumes were manually annotated by experts in the domain, to emphasize the various abdominal organs. There are nine distinct abdominal organs, namely the spleen, left kidney, right kidney, liver, gall bladder, pancreas, stomach, right adrenal and left adrenal gland. The spleen, liver, and stomach were considered the larger organs, while the kidneys, gall bladder, pancreas, and adrenal glands were considered smaller in size.

Skin lesion:
The *ISIC* dataset [25] contains dermoscopic images with pixel values $\in [0, 255]$. These images were curated by the International Skin Imaging Collaboration (ISIC) for the study of skin cancer. The training dataset consists of 2000 images with corresponding ground truth masks prepared by domain experts. The test set consists of 600 images. The value of pixel 0 in the ground truth mask corresponds to the background region, whereas the value of pixel 255 refers to the lesion region. The input images and their corresponding masks were normalized to the pixel values $\in [0, 1]$. The training data was augmented by rotation and random cropping transformation.

Cardiac organs:
The dataset *ACDC* [12] consists of 150 volumes of chest magnetic resonance images for automated cardiac diagnosis. Magnetic resonance volumes were obtained from the University Hospital of Dijon, France. The experts prepared the corresponding ground truth volumes, which report segmentation for the Right Ventricle (RV), Left Ventricle (LV), and Myocardium.

1.6.2 Software

The models presented in this thesis were implemented in Python 3.9 programming language. DL frameworks, *viz.* Tensorflow[1] and PyTorch[2], were used to develop the deep architectures. The MONAI[3] framework was used to implement data transforms and data loading for training. Python libraries like Matplotlib[4] and Seaborn[5] were used to visualize the qualitative results.

1.7 Scope and Contribution of the Thesis

This thesis aims to enhance the domain of automated medical image analysis by developing novel and resource-efficient deep learning models for the precise segmentation of anatomical structures. We present multiple computationally efficient segmentation frameworks, advancing from sophisticated attention-based Convolutional Neural Networks (CNNs) to hybrid architectures that integrate Wavelet Transforms, Vision-xLSTM, and adaptive token pruning. The generalizability of the developed models is demonstrated across different segmentation tasks, *viz.* disease-related lesions, cardiac structures and abdominal organs. Different imaging modalities, like CT, color fundus, MRI and dermoscopic images, are used to validate the robustness and adaptability of the developed models. Our objective is to develop powerful and computationally optimized architectures to address the high computational demands of state-of-the-art segmentation approaches. The ultimate goal is to provide substantial assistance to medical professionals by implementing advanced AI tools in various resource-constrained clinical environments. The efficacy of the developed architectures is demonstrated on multiple publicly available datasets, through detailed comparisons with the state-of-the-art architectures. The results of these investigations are elaborated in subsequent chapters, and illustrated in Fig. 1.11.

1.7.1 Global Context-Aware Attention for Efficient Segmentation [30], [37], [39]

Automated segmentation of anatomical structures is helpful in a wide range of modern clinical applications, ranging from early disease detection to preoperative planning. However, the variable shape, size, and structure of target regions, such as diffuse COVID-19 lesions, minute red lesions in Diabetic Retinopathy, and complex organs in abdominal CT, pose significant challenges to the generalizability of automated approaches. Standard deep models struggle to effectively distinguish between complex structures. This inspires the development of advanced frameworks to learn a richer set of features from input images.

[1] https://www.tensorflow.org/
[2] https://pytorch.org/
[3] https://monai.io/
[4] https://matplotlib.org/
[5] https://seaborn.pydata.org/

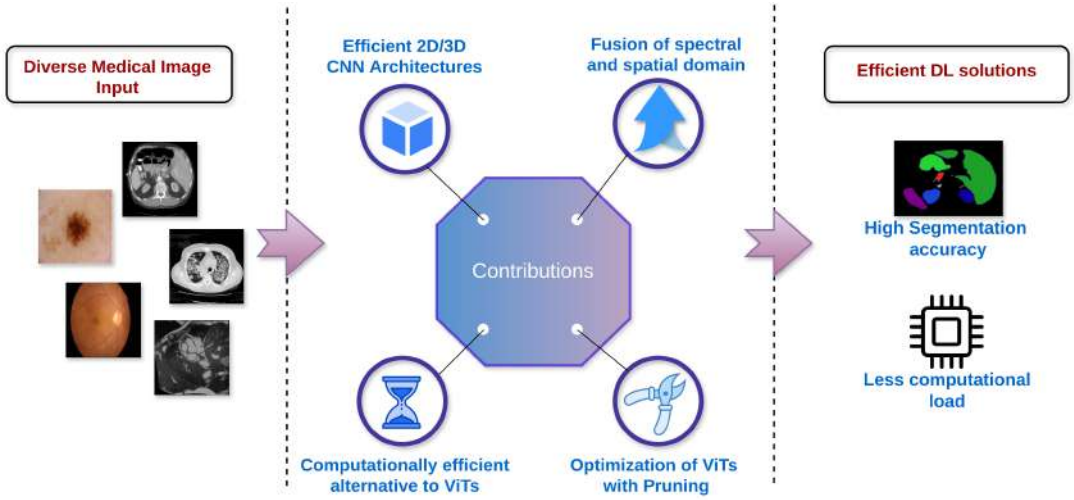

FIGURE 1.11: Schematic diagram illustrating the different research objectives addressed in chapters.

Chapter 2 details the development of a novel global context-aware attention framework, called the *Full-Scale Deeply Supervised Attention Network (FuDSA-Net)*, which aggregates multi-scalar features to create an informed attention signal. The architecture demonstrates superior performance, compared to other baselines, in segmenting COVID-19 and Diabetic Retinopathy lesions. Subsequently, the concept is evolved to the *Volumetric global Context integrated Attention Network (VoCANet)*, a 3D framework for segmenting target structures from volumetric images. The 3D framework is designed to have a lower computational complexity, without any degradation in segmentation performance. *VoCANet* outperforms state-of-the-art approaches in delineating multiple complex organs and adrenal gland tumors from abdominal CT scans, with comparatively lower computational complexity. The qualitative and quantitative results suggest that these models provide a robust and computationally efficient solution for medical image segmentation.

1.7.2 Wavelet-Enhanced Hybrid Transformer for Robust Segmentation [40]

Medical image segmentation has progressed from purely convolutional networks to hybrid architectures leveraging the benefits of CNNs and Vision Transformers. Although this fusion led to significant improvement in segmentation performance, the substantial computation demand of self-attention mechanism within ViT restricts its practical deployment in resource-constrained environments. Chapter 3 addresses these challenges by presenting a novel segmentation framework that prioritizes high segmentation accuracy in a computationally efficient manner. The *Wavelet-infused Convolutional Transformer (WaveCoformer)* combines spatial and spectral domain information for efficient segmentation. A Spectral feature Convolution *(SpectraConv)* module is introduced to

learn finer textural patterns from the wavelet decomposition of the input. Simultaneously, a Dual-Attention module filters relevant feature channels while modeling long-range global dependencies in the spatial domain. A Cross-Context Attention block integrates the information learned from the spatial and spectral domains to generate a comprehensive and robust feature representation of the target anatomical structure. *WaveCoformer* is found to surpass the segmentation accuracy of several state-of-the-art approaches in delineating abdominal organs and adrenal gland tumors, along with a lower parameter count. Extensive qualitative and quantitative evaluation on public datasets establishes the efficacy of the novel architecture.

1.7.3 Optimizing Segmentation with Vision Extended LSTM [35], [38]

The computational bottleneck associated with the quadratic computational complexity of ViTs poses a hurdle to their adaptation in resource-constrained clinical environments. Architectures with high computational demands increase processing times and necessitate sophisticated platforms for execution. Addressing this challenge is important in developing scalable AI solutions for automated medical image analysis. Chapter 4 details the development of two novel segmentation models, based on this principle. First, the novel *U-VixLSTM* is introduced to validate the superiority of Vision-xLSTMs over ViTs in medical image segmentation. The architecture achieves competitive segmentation performance, especially in precise boundary delineation with high inference speed as compared to ViT-based baselines. Next, its advanced version *Rotational U-Vision-xLSTM (Rot-UViL)* is presented. The architecture incorporates a novel *Rotational Attention Module (RAM)* to model cross-dimensional dependencies within volumetric input data. This approach is found to achieve high segmentation accuracy with computational efficiency, particularly with a lower memory footprint on the publicly available dataset.

1.7.4 Spatially-Aware Token Processing in Efficient Vision Transformers [36]

ViTs, despite providing powerful medical image segmentation models, face challenges during deployment due to their high computational demands. The self-attention mechanism often processes semantically redundant tokens, consuming significant computational resources. Algorithmic optimization techniques, such as dynamic pruning, are a promising direction for reducing computational load by selectively processing relevant tokens.

Chapter 5 outlines the Prompt-driven Adaptive Token pruning (*PrATo*) framework, which achieves runtime efficiency by dynamically adapting to the input. It uses spatial priors to guide the retention of useful tokens in ViT. The parameter-free entropy-based token relevance scoring mechanism helps identify useful tokens without additional parameter overhead. The developed token

pruning framework is incorporated into state-of-the-art ViT-based segmentation models, offering a reduction of 35-55% tokens; thus significantly reducing computational costs relative to baselines.

Chapter 2

Global Context-Aware Attention for Efficient Segmentation

" The world is full of obvious things which nobody by any chance ever observes."

— Arthur Conan Doyle, *The Hound of the Baskervilles*

2.1 Introduction

The variability of anatomical structures in medical images, related to their shape, size, and texture, poses challenges for deep learning models in achieving precise segmentation. The $U-$Net framework, although powerful, tends to propagate noisy features via the skip connections to the decoder arm; thereby affecting the segmentation quality. The popular attention mechanisms in the literature [97] do not leverage the rich multi-scalar information, which is captured hierarchically across the encoder arm.

This chapter addresses this gap by developing a novel global context-aware attention mechanism. The core idea is to aggregate feature maps from the preceding stages of the encoder arm at a particular level, before combining them with the feature maps at the corresponding stage of the decoder. This approach enables the model to develop a holistic understanding of the target structure by combining fine-grained spatial details with coarse semantic details. A novel attention module refines the aggregated multi-level features to determine relevant feature maps while highlighting informative spatial regions within them.

The *Full-Scale Deeply Supervised Attention Network (FuDSA-Net)* [39], [30] is the initial 2D model developed on the principle above. It effectively segments COVID-19 lesions from lung CT, as well as minute red lesions indicative of Diabetic Retinopathy (DR) from the colour fundus images of the eyes. The COVID-19-infected regions exhibit asymmetric shapes and positioning of affected tissues, characterized by low contrast and indistinct edges. The tiny size of the red lesions makes them difficult to identify and leads to a severe class imbalance. The network is designed to tackle the unique challenges associated with different types of lesions in these tasks.

Building on the success of *FuDSA-Net*, the model was evolved into a computationally efficient 3D framework, *Volumetric global Context integrated Attention Network (VoCANet)* [37], for segmenting target regions from volumetric medical images. It could tackle the complexities associated with segmenting multiple organs and adrenal gland tumors from abdominal CT scans. The research contributions are summarized below.

- Development of a novel global context-aware attention module which integrates multi-scalar features to effectively segment anatomical structures of diverse shapes and sizes.
- Design of 2D *FuDSA-Net*, incorporating the novel attention module to effectively segment COVID-19 and DR lesions.
- Extension of the framework to the 3D *VoCANet*, which exhibits superior segmentation performance and generalizability in multi organ and tumor segmentation tasks from volumetric abdominal CT scans.

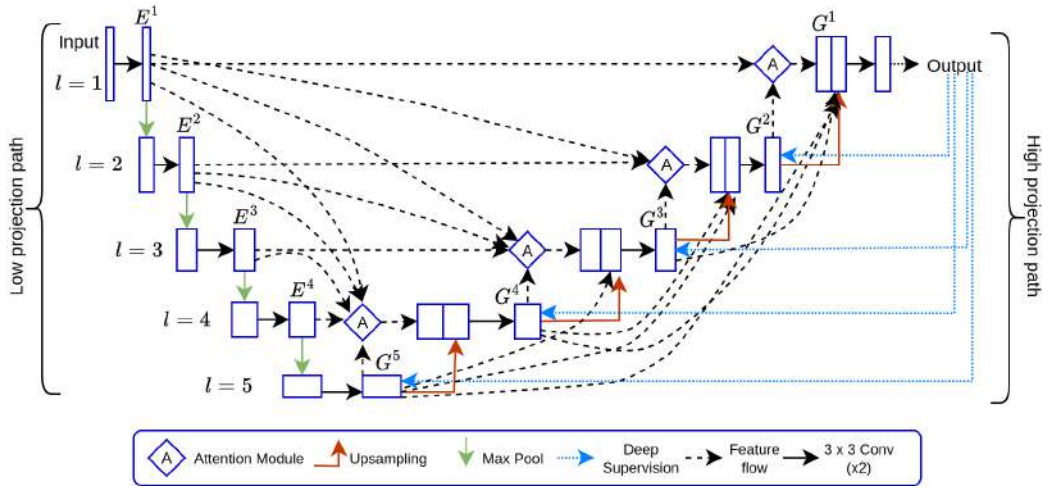

FIGURE 2.1: Architectural framework of *FuDSA-Net*.

- Improved computational efficiency by using fewer computational resources.

The rest of the chapter is organized as follows. Section 2.2 details the *FuDSA-Net* architecture, with empirical results to validate its effectiveness. Subsequently, Section 2.3 describes the evolution of this framework in the volumetric domain with *VoCANet* by presenting the architecture with performance in volumetric medical images. Finally, Section 2.4 concludes the chapter.

2.2 Full-Scale Deeply Supervised Attention Net

This section provides a comprehensive overview of the architectural features of the novel Full-Scale Deeply Supervised Attention Net *(FuDSA-Net)*, while elaborating on its different components, including the proposed attention module. Fig. 2.1 shows a schematic representation of the architectural design of *FuDSA-Net*.

This is a fully convolutional neural network, consisting of low- and high-projection paths. The objective of the low-projection path is to systematically construct high-level semantic features, with reduced resolution, from the input data at higher resolution. The high-projection path is responsible for incrementally projecting the low-resolution feature maps, obtained earlier, onto a higher-resolution segmentation output. Each stage l, along both low- and high-projection paths, consists of two convolutional layers (3×3 kernels) that are responsible for extracting meaningful information from the input acquired in the preceding stage.

The output of each stage l along the low-projection path (E^l) undergoes a spatial reduction of a factor of 2, using the Max-Pooling operation. The process of spatial reduction allows for compression with improved computational efficiency of feature volumes. Reduced feature maps need fewer parameters and computations, resulting in faster training and inference processes. This also facilitates the expansion of the receptive field; thereby, enabling the model to effectively capture a greater amount of contextual information [85].

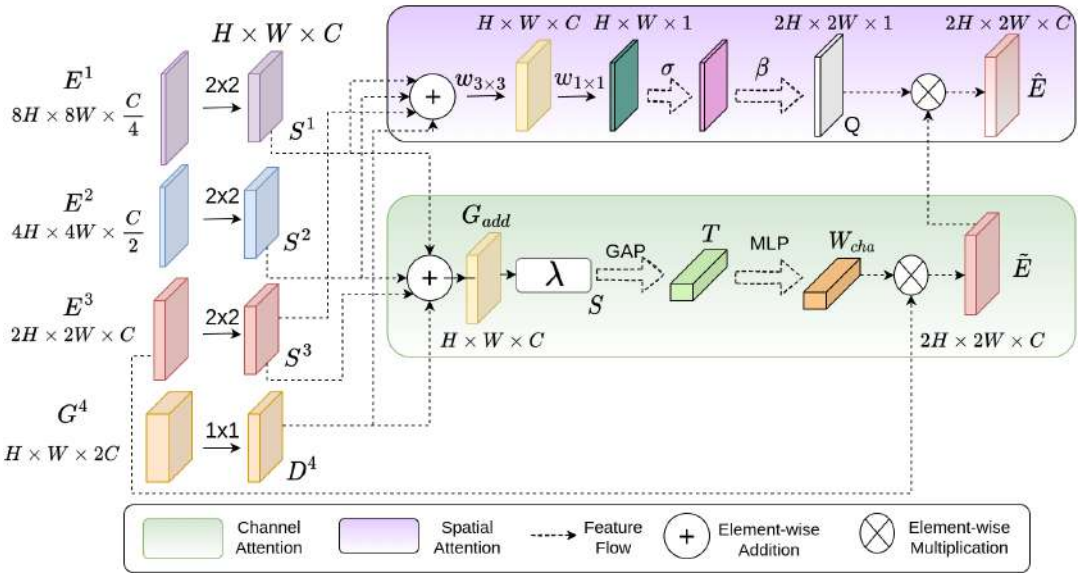

FIGURE 2.2: Attention module producing weighted feature representation $\hat{E}$ from low-projection path at stage 3.

The high-projection path $[G^5, G^4, \ldots, G^1]$ consists of a series of densely interconnected 2D convolution kernels. It enables the incorporation of a wider range of features from multiple stages. This allows the model to utilize a more extensive and diverse set of information when making final predictions. Multi-stage supervision is also incorporated to ensure improved gradient flow for overcoming vanishing gradients, while learning the more discriminative features (specific to the regions of interest) at every stage of the model. A novel attention mechanism A is incorporated in each stage of the architecture to prioritize relevant regions within the activation map volumes. This results in efficient utilization of computational resources.

2.2.1 Attention mechanism

Introducing skip connections at each stage between the low-projection and high-projection pathways facilitate the merging of detailed spatial information with more abstract, higher-level features. This integration enables the network to preserve important details that may be lost during downsampling. Nevertheless, it is frequently observed that unnecessary representations tend to persist in the initial stages, leading to a lack of precision in the segmentation output [97]. To overcome this issue, we introduce a novel attention module A at every stage of the network. This module serves to enhance the input feature volumes, enabling the model to focus on processing the most informative components of the input data. Consequently, the model optimizes computational resources by prioritizing the allocation of processing power to significant regions while minimizing processing load on the less pertinent areas.

The schematic diagram of our attention module is shown in Fig. 2.2. It incorporates both spatial and channel attention mechanisms to prioritize activation maps within the entire volume and subsequently identifies relevant

regions within those maps. Instead of exclusively refining feature volumes originating from the same stage, the module incorporates feature volumes from the preceding stages up to the current stage. This integration efficiently combines contextual information across various scales, resulting in enhanced identification and characterization of target structures that involve diverse shapes and dimensions. Let the input feature volume to the attention module at level l be represented by the feature set

$$\{E^i, G^{l+1} | i \in \{1, \ldots, l\}, E^i \in \mathbb{R}^{\frac{C}{2^{l-i}} \times 2^{(l+1)-i} H}, G^{l+1} \in \mathbb{R}^{2C \times H \times W}\},$$

where C, H and W are the channel, height and width dimensions respectively. Here $\{E^i | i \in \{1, \ldots, l\}\}$ are the feature volumes from stage 1 to l of the low-projection path while G^{l+1} is the input volume from stage $l+1$ of the high-projection path. The 2×2 convolutions are employed on the set of feature volumes $\{E^i | i \in \{1, \ldots, l\}\}$, characterized by varying height, width and channel dimensions, to achieve the desired outcome of output volumes $\{S^i | i \in \{1, \ldots, l\}, S^i \in \mathbb{R}^{H \times W \times C}\}$ having uniform dimensions concerning their height, width, depth, and channel attributes. The 1×1 convolution is applied to G^{l+1}, resulting in $D^{l+1} \in \mathbb{R}^{H \times W \times C}$ with channel reduction from $2C$ to C.

Channel attention

The lower branch of Fig. 2.2 depicts the channel attention mechanism employed to emphasize the pertinent activation maps within the input volume E^l at a particular stage l. The weights for recalibrating E^l are calculated by utilizing all the multi-scalar contextual information, captured across the feature volumes from stage 1 to l. Feature map volumes $S^1, S^2, \ldots, S^l$ and De^{l+1} are added element-wise to obtain $G_{add} \in \mathbb{R}^{H \times W \times C}$, which is mathematically expressed as

$$G_{hwc} = (\sum_{i=1}^{l} S^i_{hwc}) + D^{l+1}_{hwc}. \tag{2.1}$$

Here G_{chw} is the element at hth column, wth row and cth channel of resultant tensor G_{add}. Subsequently, the resultant tensor is subjected to hierarchically stacked 2D atrous convolutions (λ), having increasing dilation rates (r) of 2, 3 and 5 to yield output tensor $\mathbf{S} \in \mathbb{R}^{H \times W \times C}$. This enables capturing features at multiple levels of detail without a significant increase in the number of trainable parameters. The property is especially beneficial when dealing with multi-dimensional medical data, as it allows the network to effectively capture extensive relationships and contextual information without imposing substantial computational demands. Mathematically, it can be expressed as

$$\mathbf{S} = \lambda_{r=2}(G_{add}) + \lambda_{r=3}(\lambda_{r=2}(G_{add})) + \lambda_{r=5}\{\lambda_{r=3}(\lambda_{r=2}(G_{add}))\} + w_{1\times1}(G_{add}). \tag{2.2}$$

Here $w_{1\times1}$ represents point-wise convolution applied on G_{add} to mitigate the information loss resulting from the presence of gaps in atrous convolution kernels. The Global Average Pooling (GAP) aggregates the entire volumetric information across the height and width dimensions of $\mathbf{S}$ to a tensor $T \in \mathbb{R}^{C \times (HW)}$.

The output is subsequently presented to a multi-layer perceptron (MLP) to generate the final weight tensor $W_{cha} \in \mathbb{R}^{C\times1\times1}$. This is finally element-wise multiplied ($\otimes$) with E^l, to obtain the recalibrated volume $\tilde{E}^l \in \mathbb{R}^{2H\times2W\times C}$. Therefore, we have

$$W_{cha} = \sigma\{MLP(\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\mathbf{S}_{Cij})\}, \tag{2.3}$$

$$\tilde{E}^l = E^l \otimes W_{cha}, \tag{2.4}$$

where σ is the sigmoid activation function used to scale the weights within the interval $[0, 1]$.

Spatial attention

The spatial attention mechanism is depicted in the upper branch of Fig. 2.2. It is responsible for identifying the pertinent regions within the output volume of channel attention, $\tilde{E}^l$. Here $\{S^1, S^2, \ldots, D^{l+1}\}$ are (element-wise) added to efficiently combine the multi-scalar contextual information obtained from different scales along the low- and high-projection paths. This is followed by a 3×3 convolution operation ($w_{3\times3}$), to further learn non-linearity and inherent complexities. Subsequently, the resultant tensor is convolved with 1×1 kernel ($w_{1\times1}$), followed by a σ activation and bilinear upsampling operation β to generate the weight map $Q \in \mathbb{R}^{2H\times2W\times1}$ as

$$Q = \sigma[\beta\{w_{1\times1}\{w_{3\times3}(\sum_{i=1}^{l+1}S^i)\}\}]. \tag{2.5}$$

The final output volume $\hat{E}^l \in \mathbb{R}^{2H\times2W\times C}$ of the attention module is generated by element-wise multiplication of eqn. (2.5) with $\tilde{E}^l$, and expressed as

$$\hat{E}^l = \tilde{E}^l \otimes Q. \tag{2.6}$$

2.2.2 Implementation details

This section presents the datasets and preprocessing steps used to evaluate the generalizability of *FuDSA-Net*, along with the loss functions for training.

Datasets

FuDSA-Net was evaluated for COVID-19 segmentation on a combination of four COVID-19 datasets and for red lesion segmentation on IDRID and Messidor datasets. The details of the individual data are provided in Sec. 1.6.1.

All lung CT slices from the four COVID-19 datasets were resized to a dimension of 512×512. The voxel intensities of all CT volumes, from the four data sources, were clipped to limit them in the range [-1000 HU, 170 HU]. Intensity clipping eliminated undesired artifacts and noise, thereby enhancing the visualization of pertinent structures within the region of interest. In addition,

it standardized the intensity scale in various CT volumes. This was followed by intensity normalization across the resultant multi-source database. Since all CT slices in the volume did not contain COVID lesions, we selected only those having embedded lesions for training. Some datasets did not include labels for all potential COVID-19 pathologies. To address this, we combined the different pathologies from datasets 2 and 3 to create a single class representing COVID-19 lesions. The annotated samples from the four multisource datasets were combined into a single database. The combined data set was then randomly divided into five parts for a five-fold cross-validation. Combining different data sources improved the ability of the model to recognize COVID-19 lesions of varying severity.

The green channel of the RGB fundus image provides the most relevant information for segmenting red lesions [115]. Thus, this color plane was enhanced by CLAHE. Overlapping patches of size 128×128 were extracted from each enhanced image in the training dataset to reduce redundancy. Non-overlapping 128×128 patches were considered from each image for generating the test dataset.

Loss function

Independent loss functions were calculated across the five stages of the high-projection path, *viz.* $l_1, l_2, \ldots, l_5$, to provide multi-stage supervision for producing the final segmentation output map. The total loss $\mathbf{L}$ is the aggregate of individual loss, which was calculated by assigning weights ($\{\mu_i | i \in 1, \ldots, 5\}$) to each loss component. We have

$$\mathbf{L}(\{\Gamma, \hat{\Gamma}\}; \Omega) = \sum_{i=1}^{5} \mu_i l_i(\{\Gamma, \hat{\Gamma}\}; \omega_i) + \rho(\Omega), \tag{2.7}$$

where $\hat{\Gamma}$ and Γ denote the prediction and their corresponding ground truth, and Ω is the weight of the entire network. Here $\{\omega_i | i \in \{1, 2, \ldots, 5\}\}$ are the weights of the segmentation heads at the five stages of the high-projection path, and ρ denotes L_2 regularization.

The $l_1, l_2, \ldots, l_5$ denote the Focal Tversky loss (FTL) [1], an enhanced version of Tversky loss (TL) [113], to address class imbalance while focusing on hard-to-segment areas. It incorporates an additional parameter γ with the TL, to exponentially increase the loss for pixels with low predicted probability. This is expressed as

$$\begin{aligned} l_i &= \sum_{\kappa=1}^{K} FTL_\kappa = \sum_{\kappa=1}^{K} (TL)_\kappa^\gamma = \sum_{\kappa=1}^{K} (1 - TI)_\kappa^\gamma \\ &= K - \sum_{\kappa=1}^{K} \left(\frac{\sum_{i=1}^{N} \hat{y}_{\kappa,i} y_{\kappa,i} + \phi}{\sum_{i=1}^{N} \hat{y}_{\kappa,i} y_{\kappa,i} \{1 - (\alpha + \beta)\} + \theta + \phi} \right)^\gamma, \end{aligned} \tag{2.8}$$

where $\theta = \alpha \sum_{i=1}^{N} y_{\kappa,i} + \beta \sum_{i=1}^{N} \hat{y}_{\kappa,i}$, TI is the Tversky Index, $\hat{y}_{\kappa,i}$ and $y_{\kappa,i}$ are the predicted and actual ground truth values respectively for the ith pixel w.r.t.

TABLE 2.1: Ablation of *FuDSA-Net* on COVID-19 dataset.

Model	DSC	Recall	IoU
FuDSA-Net	**0.7924**	**0.8104**	**0.6681**
FuDSA-Net-I	0.7424	0.7046	0.6074
FuDSA-Net-II	0.7639	0.7246	0.6308
FuDSA-Net-III	0.7905	0.7709	0.6650

class κ, N is the total number of pixels in the input, and K is the total number of classes. The α and β are weights to control the trade-off between FP and FN, respectively, with ϕ being the additive smoothing parameter to prevent any division-by-zero error.

2.2.3 Results and discussion

Here we provide quantitative and qualitative studies evaluating the performance of FuDSA-Net.

COVID-19 images

Ablation studies involving three variants of *FuDSA-Net* were performed to examine the role of the constituent components. These are (i) *FuDSA-Net-I*, with only the spatial attention branch; (ii) *FuDSA-Net-II*, without deep supervision; (iii) *FuDSA-Net-III*, excluding the residual connections between the various stages of the decoder and (iv) *FuDSA-Net* in its full configuration. The experimental results for each of these variants are summarized in Table 2.1 for COVID-19 data. The best results are marked in bold in the table. It is observed that incorporating channel attention into *FuDSA-Net* significantly improves its performance over *FuDSA-Net-I*, as quantified by a significant gain in DSC [eq. (1.21)]. This suggests that an enhanced attention mechanism is necessary to capture the complicated lesion regions of COVID-19. Improvement is also observed through the involvement of deep supervision as well as the incorporation of intra-stage connections in decoder arms.

TABLE 2.2: Comparison of *FuDSA-Net* with baseline models on the combined COVID-19 dataset.

Model	DSC	Recall	IoU
FuDSA-Net	**0.78**	**0.79**	**0.65**
U-Net	0.63	0.59	0.48
U-Net++	0.68	0.7	0.54
Attention *U*-Net	0.65	0.64	0.52
Residual *U*-Net	0.65	0.6	0.51
U-Net 3+	0.71	0.7	0.58

FuDSA-Net was compared with $U-$ Net, $U-$ Net++, Attention $U-$ Net, Residual $U-$ Net and $U-$ Net 3 + for the COVID-19 lesion segmentation task, as summarized in Table 2.2. It is observed that *FuDSA-Net* outperforms the rest by a significant margin of approximately 7% in terms of DSC (compared

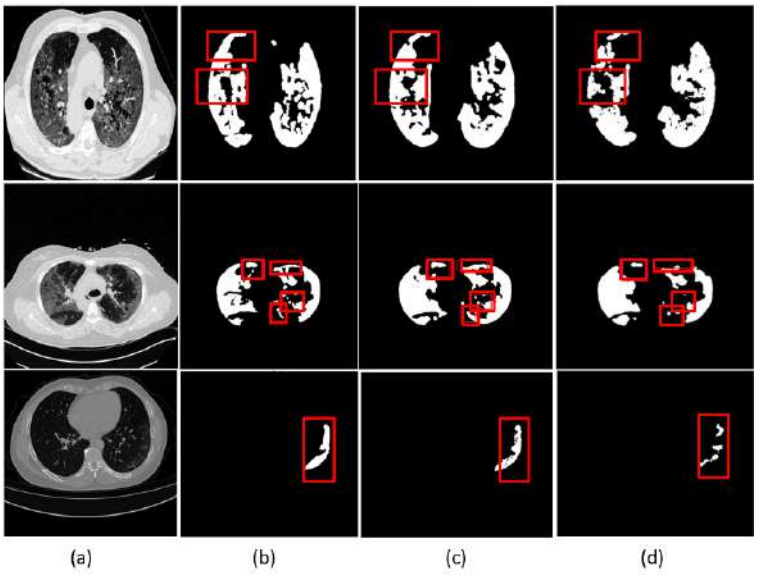

FIGURE 2.3: (a) Sample CT scan of COVID-19 affected patient, with (b) corresponding ground truth, along with predictions made by (c) proposed *FuDSA-Net*, and (d) *U*-Net++. The red box highlights the comparison area in each case.

to scores of the second best performing model, $U-$ Net 3+). A *Recall* [eq. (1.19)] of 79% signifies fewer numbers of FN pixels in the generated output. It is a critical requirement for automated segmentation methods where missing a region of the lesion might be of significant concern. A noticeable gain is also evident in terms of the value of IoU [eq. (1.20)] by our *FuDSA-Net*. This significant performance gain, both in terms of greater overlap with ground truth and stability, quantitatively validates the effectiveness of *FuDSA-Net* compared to baseline architectures.

Fig. 2.3 illustrates the qualitative results of *FuDSA-Net* and *U*-Net++ on three test cases from the COVID-19 dataset with varying levels of severity. The sample outputs of *FuDSA-Net* are comparatively closer to ground truth than the baseline *U*-Net++. In the first two rows, showing severe cases of infection, *FuDSA-Net* accurately delineates the complex boundaries of the lesions. In contrast, *U*-Net++ under-segments the areas marked by red boxes. *FuDSA-Net* exhibits a higher accuracy than *U*-Net++ in the second row with multiple scattered lesions. In the third case, corresponding to early stage infection characterized by small diffuse lesions, *U*-Net++ produces a fragmented output map. However, *FuDSA-Net*, in general, excels at generating accurate segmentation maps.

Fig. 2.4 visualizes the feature maps generated from the skip connections to study the impact of the dual attention mechanism in generating the final output in *FuDSA-Net*. The sample input maps to the dual attention block, as illustrated in Fig. 2.4(c), show irrelevant anatomical structures. However, after being processed by the attention module, the activations in the output [Fig. 2.4(d)] are focused on the true lesion areas, while activations from background

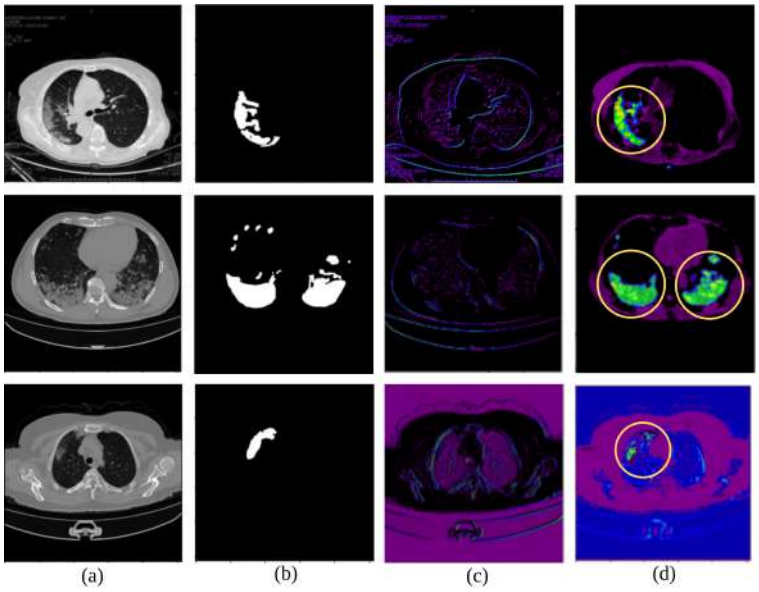

FIGURE 2.4: (a) Sample CT scan of a patient affected by COVID-19, (b) corresponding ground truth, (c) input feature maps to the attention module, and (d) output from the attention module.

regions are suppressed. This visual representation demonstrates the effectiveness of the proposed module in producing superior segmentation results.

Diabetic retinopathy

TABLE 2.3: Comparison of *FuDSA-Net* with baseline models on DR data.

Dataset	**Model**	**Metrics**		
		Recall	**Specificity**	**DSC**
Messidor	*U*-Net	0.5968	0.9811	0.5456
	U-Net++	0.5671	0.9741	0.4895
	Attention *U*-Net	0.6722	0.9767	0.5789
	U-Net 3+	0.6757	0.9445	0.6384
	Residual *U*-Net	0.6112	0.9558	0.6432
	FuDSA-Net	**0.8231**	**0.9807**	**0.6977**
IDRID	*U*-Net	0.6659	0.9075	0.6457
	U-Net++	0.4984	0.9428	0.5585
	Attention *U*-Net	0.6499	0.9117	0.6375
	U-Net 3+	0.5457	**0.9475**	0.6228
	Residual *U*-Net	0.5151	0.9408	0.5896
	FuDSA-Net	**0.6985**	0.8992	**0.6768**

Table 2.3 outlines the experimental findings comparing *FuDSA-Net* with four notable architectures, *viz.* *U*-Net, *U*-Net++, Residual *U*−Net, *U*-Net 3+ and Attention *U*-Net on the DR datasets. Quantitative findings demonstrate the consistent superiority of *FuDSA-Net* in detecting the key markers of Diabetic Retinopathy. *FuDSA-Net* surpasses all baseline models in the performance measures presented for the Messidor dataset. It achieved the highest

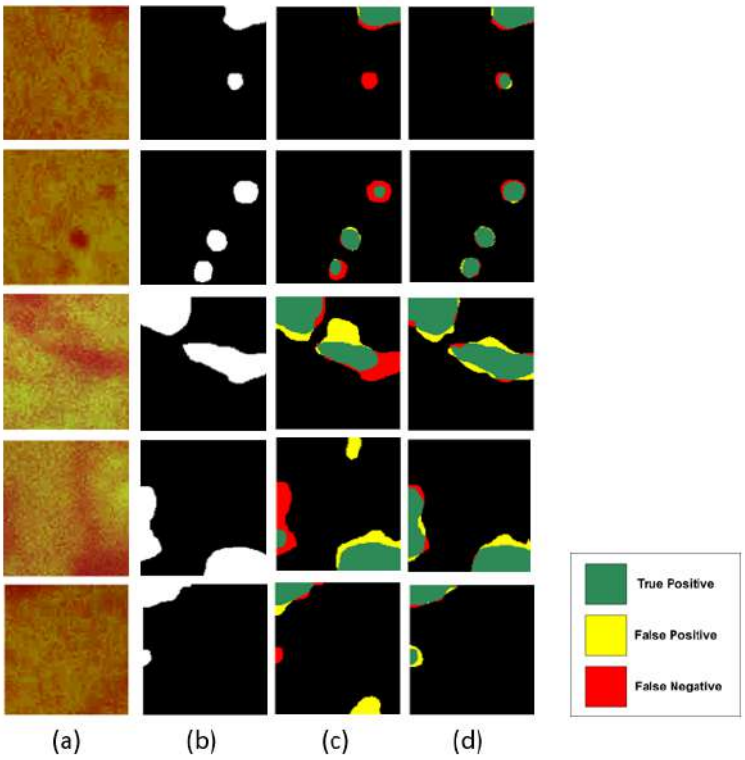

FIGURE 2.5: (a) Sample eye fundus images of DR affected patient, with (b) corresponding ground truth, along with predictions made by (c) Attention $U-$Net and (d) *FuDSA-Net*.

DSC value of 69.77%, nearly 6% improvement over the second-best performing model. The elevated *Recall* [eq: 1.19] along with the high $Specificity$ indicates that the model is proficient in identifying target regions while accurately ignoring the background. The *FuDSA-Net* exhibits a similar robust performance in the IDRID dataset as well. The model secures the highest *DSC* and *Recall* value, indicating its superior ability to detect TP pixels compared to other methods. The consistent superiority exhibited in both DR datasets validates the efficacy of the comprehensive attention mechanism in tackling the particular challenges of red lesion segmentation. This encompasses the small size of the microaneurysms and their visual similarity to other retinal features.

Fig. 2.5 illustrates a comparative analysis between the qualitative outputs of *FuDSA-Net* and Attention $U-$Net on the eye fundus images. Our model exhibits high precision in detecting minute red lesions, which are early indicators of Diabetic Retinopathy. The Attention $U-$Net either completely fails (row 1) or undersegments (rows 2 and 3) the target lesions, leading to a significant number of FN regions. In contrast, *FuDSA-Net* has comparatively higher TP regions in the predicted segmentation map. While Attention $U-$Net has a larger number of FP and FN regions while segmenting the irregularly shaped lesions, the segmentation maps from *FuDSA-Net* are visually more aligned to the ground truth.

In general, the qualitative results reinforce the findings from the quantitative analysis, demonstrating the effectiveness of the dual attention mechanism

that integrates multiscale features in *FuDSA-Net* to accurately delineate lesions having complex shapes and varying sizes, both in COVID-19 and DR datasets.

2.3 Volumetric global-Context integrated Attention Network

The principles of *FuDSA-Net* are evolved into the Volumetric global-Context integrated Attention Network (*VoCANet*) to process volumetric medical images. However, translating this concept to the 3D domain by converting 2D operations to 3D would be computationally expensive. A naive substitution of 2D convolution by 3D kernels would exponentially increase the trainable parameters, making it challenging to deploy the model in resource-constrained environments. This required a redesign of the building blocks so that the 3D model performed efficiently with reduced computational costs.

The fundamental difference between the basic building blocks of *FuDSA-Net* and *VoCANet* lies in the choice of convolution operation. While *FuDSA-Net* employs standard 2D convolutions, *VoCANet* uses 3D depthwise separable convolution kernels across the high- and low-projection paths. Depthwise separable convolution extracts features using a two-step process, leading to computational savings as follows.

- *Depthwise convolution*: At first, 3D convolution kernels are applied to each input channel independently. This captures the spatial information separately from each channel. In contrast, a single standard convolution filter processes all input channels simultaneously.

- *Point-wise convolution:* Next, a $1 \times 1 \times 1$ point-wise convolution is applied to the output of the previous step to model inter-channel dependencies.

Lemma 1 (Computational efficiency of depthwise separable convolution)**.** *Given a 3D convolution layer with N kernels of dimension $k \times k \times k$, the computational cost is reduced by a factor of $\frac{1}{N} + \frac{1}{k^3}$ by using a depthwise-separable convolution instead of the standard convolution operation.*

Proof. Let $X \in \mathbb{R}^{H \times W \times D \times C}$ be the volumetric input feature volume and $O \in \mathbb{R}^{H' \times W' \times D' \times N}$ the corresponding output feature volume. Here H, W, D are the height, width and depth of the volumetric input X, C is the number of input channels, and H', W', D' are the height, width and depth of O.
Cost of computing one single output value in $O = k \times k \times k \times C$.
$\therefore$ the total cost to compute N output channels becomes
$Cost_{std} = k^3 \times C \times H' \times W' \times D' \times N$.
Total cost of computing the depthwise convolution step, of 3D depthwise separable convolution, is given by
$Cost_{depth} = k^3 \times H' \times W' \times D' \times C$.
Similarly, the total cost of the point-wise convolution step is
$Cost_{point} = 1^3 \times C \times H' \times W' \times D' \times N$.

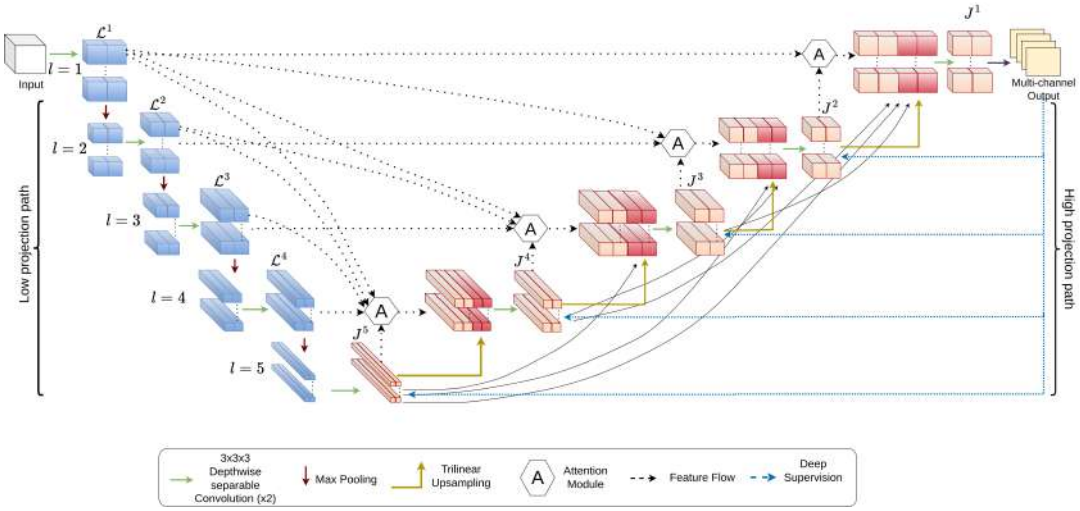

FIGURE 2.6: Architectural framework of *VoCANet*.

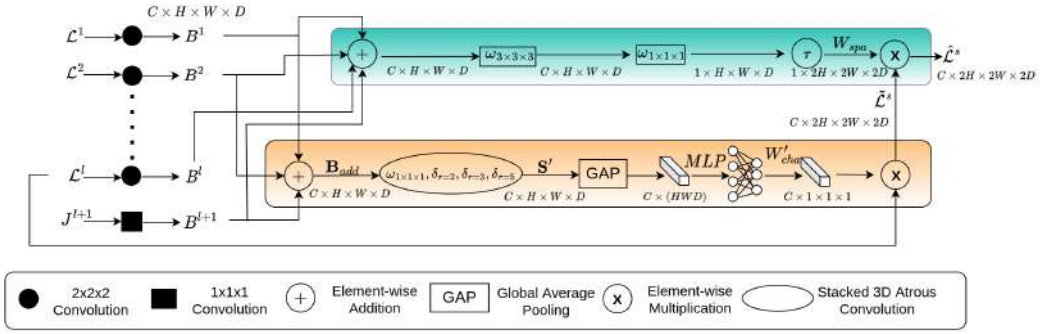

FIGURE 2.7: Attention module producing weighted feature representation $\hat{\mathcal{L}}^l$ from low-projection path at stage l.

Thus, the computational reduction ratio is expressed as

$$r = \frac{Cost_{depth} + Cost_{point}}{Cost_{std}} = \frac{H' \times W' \times D' \times C(k^3 + N)}{k^3 \times C \times H' \times W' \times D' \times N} = \frac{1}{N} + \frac{1}{k^3}. \quad (2.9)$$

□

The 3D depthwise separable convolution significantly reduces computational time by separating spatial context learning while capturing interdependencies between channels. This key adaptation allows *VoCANet* to efficiently process volumetric inputs with low computational burden. Fig. 2.6 illustrates the architectural framework of *VoCANet*.

2.3.1 Attention mechanism

The attention module of *VoCANet*, as illustrated in Fig. 2.7, is a direct 3D improvement of the 2D attention module of *FuDSA-Net*. It conforms to the core idea of refining aggregated global context information from the multiple levels of the low-projection path. Two key features of this novel attention module

that mitigate the substantial increase in computing load by substituting 2D operations with their 3D equivalents are given below.

- The channel attention branch employs hierarchically stacked 3D atrous convolutions with progressively increasing dilation rates. This allows the module to acquire spatial information with an expanding receptive field, without a substantial increase in trainable parameters.
- The sequential positioning of channel and spatial attention mechanisms is memory-efficient. Unlike alternative parallel configurations, it eliminates the need to retain the intermediate findings of each mechanism. Using the output of channel attention as the input for spatial attention results in a substantial reduction in memory consumption compared to the corresponding parallel approach.

Let the aggregated input volume to the attention module, from multiple levels at stage l, be represented as

$$\{\mathcal{L}^i, J^{l+1} | i \in \{1, \ldots, l\}, \mathcal{L}^i \in \mathbb{R}^{\frac{C}{2^{l-i}} \times 2^{(l+1)-i} H \times 2^{(l+1)-i} W \times 2^{(l+1)-i} D}, J^{l+1} \in \mathbb{R}^{2C \times H \times W \times D}\}.$$

Here, $\{\mathcal{L}^i | i \in \{1, \ldots, l\}\}$ denotes the feature volumes of stages 1 to l of the low-projection path, with varying height, width, depth, and channel dimensions, and J^{l+1} represents the input volume of stage $l+1$ of the high-projection path. The $2 \times 2 \times 2$ convolutions are applied to $\{\mathcal{L}^i | i \in \{1, \ldots, l\}\}$ to produce output volumes $\mathbf{B} = \{B^i | i \in \{1, \ldots, l\}, B^i \in \mathbb{R}^{C \times H \times W \times D}\}$ with consistent height, width, depth and channel dimensions. J^{l+1} is subjected to $1 \times 1 \times 1$ convolution, which yields $B^{l+1} \in \mathbb{R}^{C \times H \times W \times D}$ with a reduction in the channels from $2C$ to C.

The 3D **channel attention** mechanism, to suppress irrelavant channels in the 3D domain, is mathematically expressed as

$$\mathbf{B_{add}} = \sum_{i=1}^{l+1} B^i, \tag{2.10}$$

$$\mathbf{S'} = \delta_{r=2}(\mathbf{B_{add}}) + \delta_{r=3}(\delta_{r=2}(\mathbf{B_{add}})) + \delta_{r=5}\{\delta_{r=3}(\delta_{r=2}(\mathbf{B_{add}}))\} + w_{1\times1\times1}(\mathbf{B_{add}}), \tag{2.11}$$

with eqn. (2.10) depicting the aggregation of activation volumes $\{B^i | i \in \{1, \ldots, l\}$ using element-wise addition, eqn. (2.11) corresponding to the application of stacked 3D atrous convolutions δ with increasing dilation rates r to produce $\mathbf{S'} \in \mathbb{R}^{C \times H \times W \times D}$, and $w_{1\times1\times1}$ being the 3D point-wise convolution. The final generation of channel weights from $\mathbf{S'}$ is represented as

$$W'_{cha} = \sigma\{MLP(\frac{1}{H \times W \times D} \sum_{i=1}^{H} \sum_{j=1}^{W} \sum_{k=1}^{D} \mathbf{S'}_{Cijk})\}, \tag{2.12}$$

$$\tilde{\mathcal{L}}^l = \mathcal{L}^l \otimes W'_{cha}, \tag{2.13}$$

where σ is the sigmoid function to squash the channel weights within the range $[0, 1]$, and $\tilde{\mathcal{L}}^l$ is the re-calibrated volume subsequently provided as input to the spatial attention mechanism.

The 3D **spatial attention** block of the proposed attention mechanism identifies important spatial regions within $\tilde{\mathcal{L}}^l$, analogous to its 2D counterpart in *FuDSA-Net*. The spatial weight generation $W_{spa} \in \mathbb{R}^{1\times 2H\times 2W\times 2D}$ is depicted as

$$W_{spa} = \sigma[\tau\{w_{1\times 1\times 1}\{w_{3\times 3\times 3}(\sum_{i=1}^{l+1} B^i)\}\}], \tag{2.14}$$

where τ is the trilinear upsampling operation. The final output volume $\hat{\mathcal{L}}^l \in \mathbb{R}^{C\times 2H\times 2W\times 2D}$ of the 3D attention module is generated as

$$\hat{\mathcal{L}}^l = \tilde{\mathcal{L}}^l \otimes W_{spa}. \tag{2.15}$$

2.3.2 Implementation details

This section provides the datasets and their preprocessing steps used to evaluate the generalizability of *VoCANet*. It also presents the loss functions used for training.

Datasets

VoCANet was trained and validated using the *Synapse* and *AdrenalSeg* datasets. The details of the individual data are provided in Sec. 1.6.1.

The voxel intensities of the abdominal CT volumes of *Synapse* and *AdrenalSeg* were adjusted by clipping them within the range of -170 HU to 250 HU. This improved visibility of anatomical structures and eliminated any unwanted noise. Subsequently, intensity normalization was performed to rescale intensity values within the interval [0,1]. Data augmentation techniques, such as flipping along the height, width, and depth dimensions, random rotation, and introduction of random shifts in intensity values, were used to artificially expand the size of the training dataset. This helped mitigate the limited availability of data.

Loss function

VoCANet employed the same Focal Tversky loss function used to train *FuDSA-Net*, to effectively handle class imbalance while delineating the hard-to-segment region(s). Detailed formulation of the loss function is provided in Section 2.2.2.

2.3.3 Results and discussion

The variants of *VoCANet* were evaluated on *Synapse*, in an ablation study reported in Table 2.4. Four configurations of *VoCANet*, namely *VoCANet*-I (utilizing channel attention only), *VoCANet*-II (utilizing spatial attention only), *VoCANet*-III (comprising four stages), and *VoCANet*-IV (incorporating both spatial and channel attention and consisting of five stages) were compared in terms of *DSC*. Segmentation performance for challenging organs with smaller

TABLE 2.4: Ablation study on *VoCANet* for *Synapse* dataset.

Methods	DSC						
	Spleen	**Right Kidney**	**Left Kidney**	**Gall Bladder**	**Liver**	**Pancreas**	**Mean**
VoCANet-I	**0.9088**	0.9242	0.8838	0.5934	**0.9539**	0.754	0.8364
VoCANet-II	0.8979	0.904	0.7617	0.7414	0.9507	0.7134	0.8282
VoCANet-III	0.881	0.9244	0.8854	0.6975	0.9532	0.6166	0.8264
VoCANet-IV	0.9045	**0.9294**	**0.9262**	**0.7842**	0.9527	**0.7669**	**0.8773**

size and inconsistent positioning, *viz.* the pancreas and gall bladder, decreased significantly for the single attention variants (*VoCANet-I and II*). This reinforces the findings of *FuDSA-Net*, demonstrating the effectiveness of spatial and channel attention mechanisms for the best segmentation performance. Furthermore, the effect of network depth on volumetric segmentation was analyzed. It illustrates the complete five-stage network, with both attention mechanisms outperforming all other variants. This highlights the effectiveness of deeper networks in capturing intricate details, as is necessary to efficiently segment complex anatomical structures.

TABLE 2.5: Comparison of *VoCANet* with baseline models on *Synapse* and *AdrenalSeg* datasets.

Model	*Synapse*									*AdrenalSeg*			TFLOPs
	DSC							Mean IoU	Mean HD95	DSC	IoU	HD95	
	spleen	**rkid**	**lkid**	**gall**	**liver**	**pancreas**	**Mean**						
U-Net	0.9112	0.9007	0.9181	0.5645	0.9572	0.6967	0.8247	0.7455	53.46	0.6784	**0.7156**	60.22	3.587 T
V-Net	0.8874	0.9251	0.9244	0.5858	0.9487	0.7511	0.8371	0.7463	29.52	0.6283	0.5057	72.59	1.453 T
U-Net++	**0.9118**	0.9196	0.8905	0.6921	0.9524	0.7536	0.8533	**0.7587**	73.05	0.6799	0.6986	56.49	2.26 T
Attention *U*-Net	0.9109	0.877	0.872	0.5835	**0.9585**	0.5566	0.7931	0.6937	60.82	0.7055	0.57	81.82	3.616 T
U-Net with EfficientNet-b0 backbone	0.8541	0.8919	0.8804	0.637	0.9077	0.5018	0.7788	0.679	79.28	0.6521	0.5214	56.38	2.024 T
U-Net with EfficientNet-b1 backbone	0.7414	0.8059	0.8256	0.55	0.9019	0.5416	0.7277	0.5986	85.5	0.6138	0.5212	95.61	2.026 T
U-Net 3+	0.8736	0.9229	0.8584	0.6659	0.9383	0.6823	0.8236	0.7207	64.81	0.6864	0.5587	48.01	25.988 T
TransUNet	0.852	0.8828	0.785	0.7608	0.587	0.7218	0.7649	0.7072	35.65	0.7319	0.6066	35.77	3.057 T
VoCANet	0.9045	**0.9294**	**0.9262**	**0.7842**	0.9527	**0.7669**	**0.8773**	0.7566	**14**	**0.7407**	0.6106	**34.79**	**0.665 T**

Table 2.5 presents the quantitative results from comparing *VoCANet* against different state-of-the-art architectures. *VoCANet* demonstrates the highest average *DSC* of 87.73% and 74.07% in Table 2.5, compared to other methods. Our model demonstrates superior performance in segmenting smaller organs, *viz.,* gall bladder and pancreas. However, $U-$Net$++$ and Attention $U-$Net had the highest *DSC* in accurately delineating the spleen and liver. These were 0.73% and 0.58% higher than *VoCANet*, respectively. The highest average *DSC* obtained by *VoCANet* (in both cases) suggests that the predicted segmentation map demonstrates a high level of concordance with the ground truth; indicating precise and efficient segmentation. *VoCANet* also outperforms other methods in precisely delineating the boundaries by scoring the lowest mean $HD95$. The efficacy of *VoCANet* remained consistent in the Adrenal tumor segmentation task, as evidenced by the highest *DSC* and lowest $HD95$ metric values compared to other baselines in Table 2.5. Achieving the highest *DSC* and lowest *HD95* values on both the segmentation tasks is a reliable indicator of the model's robustness; demonstrating its consistent performance across different datasets.

VoCANet has the lowest $TFLOPs$ compared to other state-of-the-art methods. This shows that *VoCANet* outperforms comparable segmentation models in terms of inference speed, energy consumption, and adaptability to low-resource

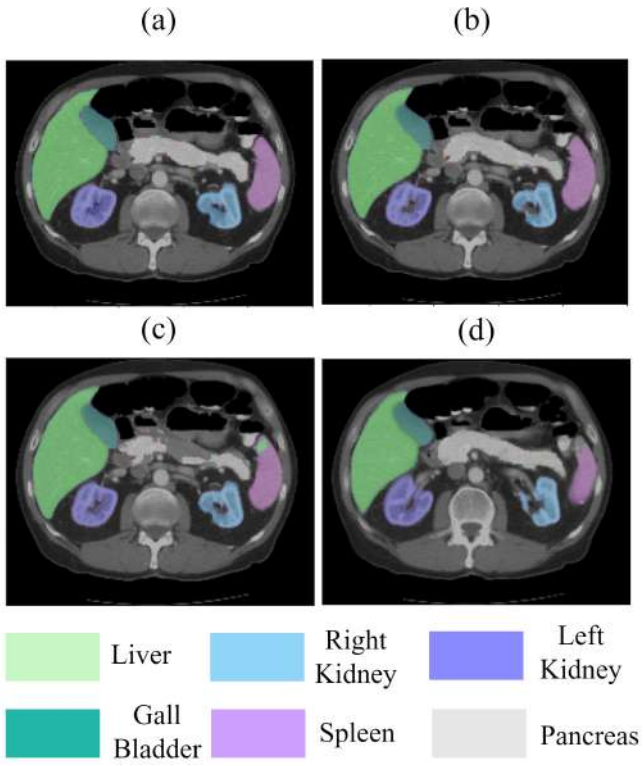

FIGURE 2.8: (a) Sample abdominal CT slice with ground truth, along with predictions made by (b) *VoCANet* (c) Attention $U-$Net, and (d) $V-$Net.

devices. Reduced computing needs lead to lower energy consumption, resulting in a more sustainable solution with a reduced carbon footprint [31].

Figs. 2.8-2.9 illustrate a visual comparison of the segmentation outputs obtained from *VoCANet* and other stated approaches. The results of both datasets suggest that *VoCANet* produces segmentation maps with higher efficiency compared to the other approaches. The segmentation outputs obtained from Attention U-Net and $V-$Net in the multiorgan segmentation task, as illustrated in Fig. 2.8(c)-(d), exhibit a greater number of FN pixels while segmenting the pancreas and spleen; thereby, leading to under-segmentation of the organs. Similarly, Fig. 2.9(c) shows that the segmentation map from $U-$Net with the EffficientNet-b0 backbone produces a suboptimally segmented tumor region. In contrast, the predicted mask generated by the $V-$Net model, as depicted in Fig. 2.9(d), exhibits an oversegmented tumor region. The segmentation mask generated by the *VoCANet* model, on the other hand, shows a reduced number of FP and FN pixels. This indicates its proficiency in accurately identifying and outlining target structures or regions from input data.

The success of *FuDSA-Net* and *VoCANet*, over the baseline models, can be attributed to architectural choices for overcoming the drawbacks of standard encoder-decoder frameworks. The aggregated multi-scalar features provide richer information to the attention module, compared to the standard local skip connections in $U-$Net and Attention $U-$Net. The dual attention mechanism determines the relevant features, along with their spatial location, unlike the

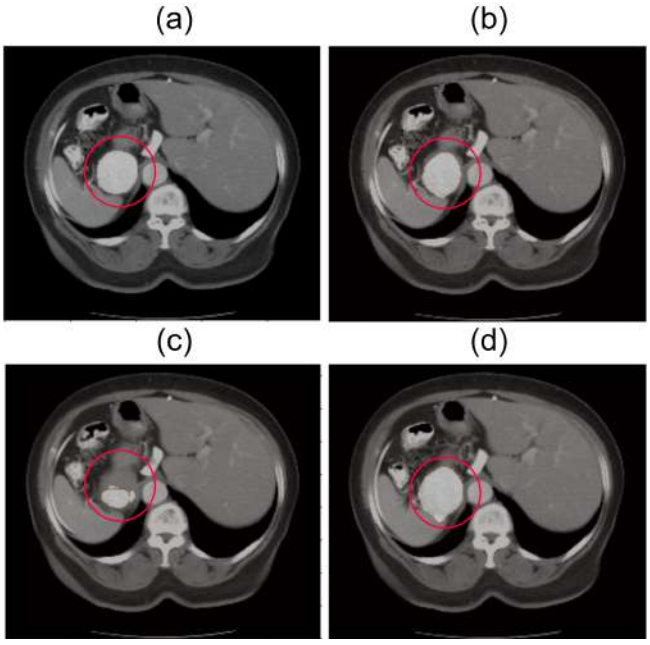

FIGURE 2.9: (a) Sample abdominal CT slice with ground truth, along with predictions made by (b) *VoCANet* (c) $U-$Net with EfficientNet-b0 backbone, and (d) $V-$Net. The red circle represents the tumor region in each output.

single spatial attention mechanism of Attention $U-$Net. In addition, the deep supervision mechanism, with the entire high-projection path, ensures strong gradient flow to mitigate vanishing gradient issues. It also enables intermediate layers to learn strong discriminative features.

2.4 Conclusion

This chapter addressed the challenge of accurately segmenting anatomical structures, having complex shapes and variable sizes. Two novel segmentation frameworks, *viz. FuDSA-Net* and *VoCANet*, were developed for handling 2D and 3D medical image datasets, respectively. *FuDSA-Net* outperformed several state-of-the-art approaches on challenging segmentation tasks related to COVID-19 and DR. It was extended to the development of *VoCANet*, a computationally efficient 3D segmentation model for volumetric segmentation tasks. The quantitative and qualitative results for both models, across different segmentation tasks, validated our hypothesis regarding leveraging and refinement of multi-scalar features for efficient and robust segmentation.

The success of the framework across different types of target structures, *viz.* lesions, organs and tumors in both 2D and 3D domains, signifies the generalizability capacity of the models. The high computational efficiency of *VoCANet* makes it a promising solution for deployment in resource-constrained environments, allowing faster and precise diagnosis.

While the developed frameworks efficiently segment target structures in different medical imaging modalities, their dependence on convolution operations limits the capture of long-range dependencies. This affects their ability to learn features of anatomical structures with variable sizes. The following chapter addresses this drawback by introducing a hybrid CNN-Transformer model, that leverages both spatial and spectral characteristics for robust segmentation performance.

Chapter 3

Wavelet-Enhanced Hybrid Transformer for Robust Segmentation

" By plucking her petals, you do not gather the beauty of the flower. "

— Rabindranath Tagore, *Stray Birds*

3.1 Introduction

Automated medical image segmentation is a significant step in modern clinical decision making, providing a crucial tool for improving patient care. Deep learning frameworks, such as CNN and ViT, are popularly used for this task because of their ability to automatically learn relevant features with minimal human intervention. Hierarchical modeling of representations from low-level textural patterns, parameter sharing, and translational invariance makes CNNs a widely preferred choice in medical imaging tasks. They accurately delineate the complex boundaries of the anatomical structures, as is necessary for high-quality segmentation. As demonstrated in the previous chapter, the integration of attention modules into $U-$Net-based segmentation frameworks improves the representational ability of the model by highlighting salient activations corresponding to target regions.

Despite being proficient in analyzing fine-grained features, the local receptive field of CNNs prevents them from acquiring a global view of the entire target structure. This adversely affects the overall performance of segmentation in complex target organs or tissues. The principle of local analysis impacts the efficacy of attention modules designed solely using convolution operations. The recalibration weight of the pixel (i, j), generated by the convolutional attention modules, is calculated using information from a small local neighbourhood around it. However, to understand the relationship between a distant pixel and the target pixel, the information needs to pass through several layers of convolution operations; thus diluting the relevant signals. This critical gap motivates exploring beyond the purely convolutional attention to effectively capture the long-range dependencies.

ViTs are a powerful alternative to address the challenge of modeling global contextual information. The self-attention mechanism of ViTs efficiently captures the relationships between different patches of the input image, allowing the model to handle global contextual information in the spatial domain. It learns the characteristics related to target objects with varying shapes and sizes. However, CNNs and ViTs exhibit a complementary nature [98]. While CNNs effectively model intricate textural details, ViTs are proficient in analysing comprehensive structural information of the target region. Incorporating both types of detail facilitates a comprehensive understanding of anatomical structures, resulting in improved segmentation accuracy.

Recently, the use of deep learning methods in conjunction with spectral domain analysis, particularly wavelets, has resulted in notable improvements in different medical image analysis tasks. Discrete Wavelet Transform (DWT) offers several advantages in the context of medical images, *viz.* multiresolution analysis and good localization of dominant frequencies in the input image [80]. This enables learning of multi-scalar features corresponding to the target structures having variable shapes.

The aforementioned advances catalyzed the development of the *Wavelet-infused Convolutional Transformer (WaveCoformer)* [40]. It learns a cohesive representation that encapsulates the textural intricacies and global context of the target structures, from both the spatial domain and wavelet coefficients. The integration of the CNN and ViT output is performed in parallel, to effectively leverage their benefits. This intertwining facilitates the simultaneous integration of textural and structural information, leading to enhanced performance. The research contributions are summarized below.

- A computationally efficient convolution block, called the Spectral feature Convolution *(SpectraConv)* module, learns textural patterns and finer low-level details from the spectral domain. It extracts information at multiple scales, to effectively model ROIs exhibiting diverse shapes and sizes.
- A dual-attention *(DA)* initially emphasizes prominent feature maps within the input volume. This is followed by the computation of a self-attention mechanism, using shifted windows to capture global dependencies among various spatial locations within the image. It allows the model to selectively focus only on the relevant maps; thereby, reducing the computational cost at the subsequent steps involving self-attention.
- A novel cross-context attention module efficiently combines the features captured by the *SpectraConv* and *DA* blocks. The integration allows the network to simultaneously incorporate local and global information, leading to a richer and more complete representation of input data.

The remaining chapter is organized as follows. Section 3.2 presents a detailed description of *WaveCoformer*, highlighting its specific characteristics and distinguishing features. The details of the implementation of the method are presented in Section 3.3. The qualitative and quantitative results obtained using the *WaveCoformer* for segmentation of Adrenocortical carcinoma (ACC) in the *AdrenalSeg* dataset and multi-organ segmentation of abdominal images on the *Synapse* dataset, are described in Section 3.4. Finally, Section 3.5 summarizes the concluding remarks.

3.2 Wavelet-infused Convolution Transformer

The *WaveCoformer* is a five-tier hybrid convolution-transformer network that involves abstraction and expansion pathways. The primary goal of the abstraction path is to systematically analyze the volumetric input data hierarchically, across multiple levels, to generate an abstract global representation of the image. The expansion path incrementally constructs the high-dimensional segmentation output, based on the low-dimensional contextual representation obtained from the preceding abstraction path. A schematic representation of the architectural layout of the model is shown in Fig. 3.1.

The Abstraction path consists of multiple intertwined convolution and transformer modules that simultaneously acquire features from both spectral and

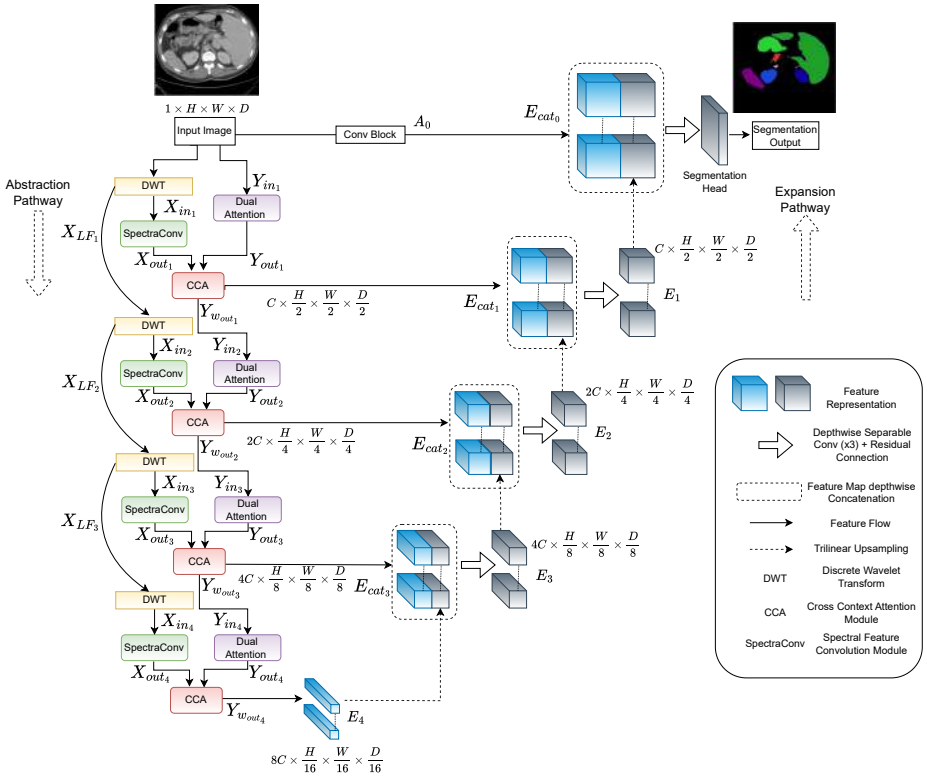

FIGURE 3.1: Architectural framework of *WaveCoformer*.

spatial domains at each stage. The *Spectral feature Convolution module (SpectraConv)* captures the intricate textural characteristics of the target region from the spectral representation of the image. On the other hand, the broader global structure information is encoded by the transformer module, named *Dual-Attention module (DA)*. The *Cross-Context Attention module (CCA)* integrates the complementary representations, acquired by the spectral convolution and transformer modules, to promote the exchange of information between them. This enables the network to take advantage of their complementary strengths along each pathway. This leads to enhanced overall performance, along with the ability to capture intricate relationships within the input image.

The output of each stage (in the Abstraction path) is directed to its corresponding counterpart at the same level of the Expansion path, through skip connections. This facilitates the integration of feature maps originating at various levels of abstraction. It also ensures a judicious combination of the finer textural details from earlier levels with the coarser semantic information of the deeper levels; thereby, resulting in enhanced context-sensitive predictions.

Every level l in the Expansion path uses a trilinear upsampling operation τ, to increase the spatial dimension of the feature maps obtained from the preceding level E_{l+1}. This helps align the spatial dimensions of the feature maps with those received from the corresponding level l from the Abstraction path $Y_{w_{out_l}}$. The upsampled volume is thus combined with the feature maps from the corresponding level of the Abstraction path, to create the fused representation

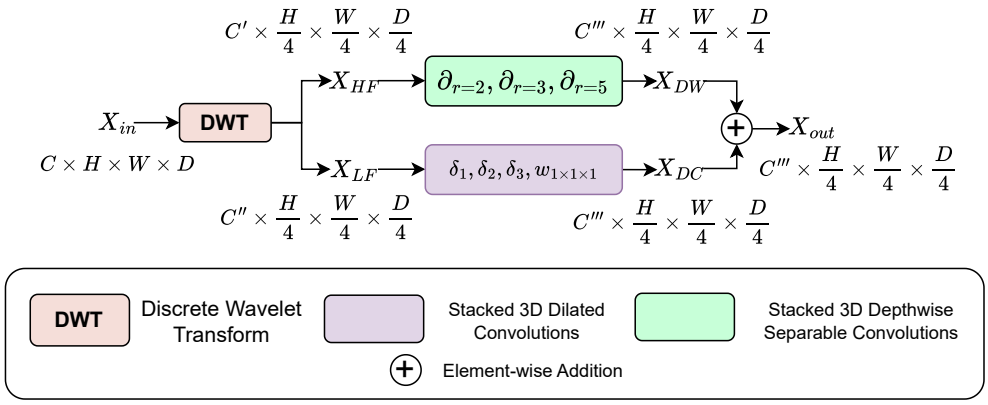

FIGURE 3.2: *SpectraConv* module for extracting textural features from low- and high-frequency components of DWT, denoted by X_{HF} and X_{HL} respectively.

E_{cat}. This is then subjected to depth-wise separable convolution $\partial_i | i \in \{1, 2, 3\}$ to produce the final output volume E_l. Residual connection is incorporated to mitigate information loss and to ensure improved gradient flow. Components of *WaveCoformer* along Abstraction path, encompassing modules *SpectraConv*, *Dual-Attention* and *Cross-Context Attention*, are elaborated below.

3.2.1 Spectral feature Convolution

Fig. 3.2 provides a diagrammatic overview of the module. It is located at different depths along the Abstraction path of *WaveCoformer*. The DWT decomposes an input image into two subbands, *viz.* the approximation subband [low-frequency component (X_{LF})] and the detailed subband [high-frequency component (X_{HF})]. Haar wavelet, with periodization at the image boundaries, was used for our experiments. Because of their simplicity, Haar wavelets provide computational efficiency. They offer enhanced delineation of target structures. Periodization alleviates boundary artifacts that could interfere with accurate boundary delineation [9]. While the low-frequency component corresponds to the overall shape and structure, the high-frequency component refers to the finer textural and edge details. The X_{LF} and X_{HF} are processed simultaneously, with the X_{LF} subjected to multiple dilated convolutions and the X_{HF} undergoing multiple depth-wise separable convolutions to extract relevant features.

Parallel processing speeds up training and inference. Dilated convolutions learn information from a broader receptive field while minimizing the parameter count. They help extract the global context from low-frequency structural details while minimizing computational overhead. Depth-wise separable convolutions, on the other hand, improve computational efficiency and memory usage, while capturing dense features similar to standard convolution kernels [22]. Thus, depth-wise separable convolutions effectively extract finer textural

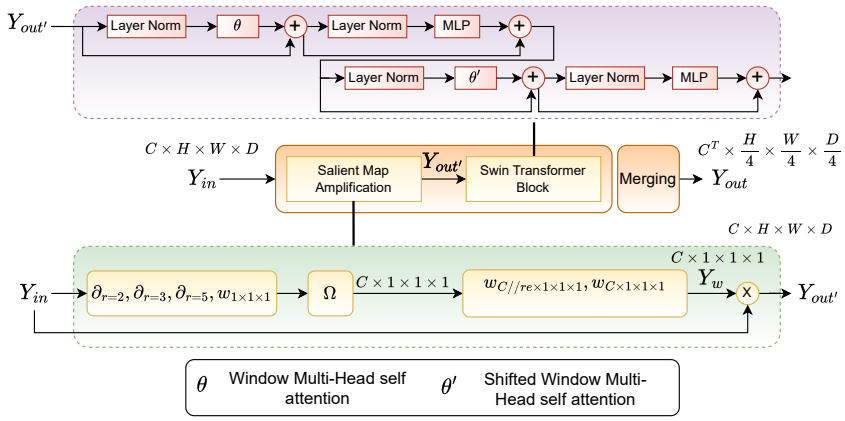

FIGURE 3.3: *Dual-Attention* module for sequentially amplifying relevant input maps.

details from high-frequency components. The *SpectraConv* module excels at acquiring relevant attributes, while simultaneously presenting a computationally efficient framework.

Let the input volume fed to *SpectraConv* be denoted as X_{in}, where $X_{in} \in \mathbb{R}^{C\times H\times W\times D}$, with C, H, W, D, indicating the number of channels, height, width and depth dimensions of X_{in}, respectively. The $X_{LF} \in \mathbb{R}^{C''\times \frac{H}{4}\times \frac{W}{4}\times \frac{D}{4}}$ undergoes a chain of hierarchically stacked 3D dilated convolutions, represented by $\delta \in \mathbb{R}^{C'''\times k\times k\times k}$. Here, the dilation rate progressively increases through r =2, 3 and 5, to produce the output volume $X_{DC} \in \mathbb{R}^{C'''\times \frac{H}{4}\times \frac{W}{4}\times \frac{D}{4}}$, with k corresponding to the size of the 3D dilated convolution kernel. Multiple dilated convolution kernels, in a stacked configuration, facilitate hierarchical enlargement of the receptive field of the module. It allows *SpectraConv* to capture information within a wider contextual scope. This is represented as

$$X_{DC} = \delta_{r=2}(X_{LF}) \oplus \delta_{r=3}\{\delta_{r=2}(X_{LF})\}\oplus \delta_{r=5}[\delta_{r=3}\{\delta_{r=2}(X_{LF})\}] \oplus w_{1\times1\times1}(X_{LF}), \quad (3.1)$$

with $w_{1\times1\times1}$ denoting the point-wise convolution on X_{LF} to alleviate any loss of detail (occurring due to the void presented in dilated convolutions), and $\oplus$ representing the element-wise addition operation. A series of 3D depth-wise separable convolution kernels $[\partial_i | i \in \{1, 2, 3\}]$ is applied on $X_{HF} \in \mathbb{R}^{C'\times \frac{H}{4}\times \frac{W}{4}\times \frac{D}{4}}$, to generate the output volume $X_{DW} \in \mathbb{R}^{C'''\times \frac{H}{4}\times \frac{W}{4}\times \frac{D}{4}}$. The final output volume $X_{out} \in \mathbb{R}^{C'''\times \frac{H}{4}\times \frac{W}{4}\times \frac{D}{4}}$ is computed as

$$X_{out} = X_{DW} \oplus X_{DC}. \quad (3.2)$$

Additionally, X_{LF} is forwarded to the next stage (or level) along the Abstraction pathway (of Fig. 3.1), for subsequent decomposition by DWT; thereby, allowing multi-resolution analysis at different degrees of granularity.

3.2.2 Dual-Attention

Fig. 3.3 illustrates the architectural details of the *Dual-Attention (DA)* module. The DA initially assigns higher weights to prominent channels in the input volume. Implementing self-attention with shifted windows (as in Ref. [52]) reduces quadratic computational complexity. Focusing on relevant activation feature maps in the input volume helps the module allocate computational power to the most informative ones. Such selective focus reduces the computational burden of subsequent self-attention steps.

Let $Y_{in} \in \mathbb{R}^{C\times H\times W\times D}$ be the input volume to the *DA* module. The Y_{in} is first subjected to multiple depth-wise separable convolutions $[\partial_i | i \in \{1,2,3\}]$, with hierarchically increasing dilation rates r =2, 3 and 5, and an $1\times 1\times 1$ convolution operation $w_{1\times 1\times 1}$. The resulting output volume $Y_1 \in \mathbb{R}^{C\times H\times W\times D}$ becomes

$$Y_1 = \partial_{r=2}(Y_{in}) \oplus \partial_{r=3}\{\partial_{r=2}(Y_{in})\} \oplus \partial_{r=5}[\partial_{r=3}\{\partial_{r=2}(Y_{in})\}] \oplus w_{1\times 1\times 1}(Y_{in}). \quad (3.3)$$

The depthwise convolution $\Psi \in \mathbb{R}^{C\times H\times W\times D}$ in Y_1 squeezes its dimension to $C\times 1\times 1\times 1$. This yields a singular value for every feature map; thereby, effectively encapsulating the statistical characteristics in terms of their depth. Subsequently, the final weight map $Y_w \in \mathbb{R}^{C\times 1\times 1\times 1}$ is evaluated as

$$Y_w = \sigma\{w_{C\times 1\times 1\times 1}[\lambda\{w_{C//r\times 1\times 1\times 1}(\Psi\overline{Y_1})\}]\}. \quad (3.4)$$

Here, $w_{C//re\times 1\times 1\times 1}$ and $w_{C\times 1\times 1\times 1}$ are the corresponding convolution kernels, with re, λ and σ representing the reduction ratio, LeakyReLU activation function, and the sigmoid activation function, respectively. A reduction ratio of 16 was taken for our implementation. The scaled feature volume $Y_{out'} \in \mathbb{R}^{C\times H\times W\times D}$ is finally obtained by element-wise multiplication ($\otimes$) of the scaling factors with Y_{in}. We have

$$Y_{out'} = Y_w \otimes Y_{in}. \quad (3.5)$$

The second phase is analogous to the self-attention mechanism with shifted windows in Ref. [52]. There are two components *viz.* Window Multi-Head self-attention (θ) and Shifted-Window Multi-Head self-attention (θ'), as indicated in Fig. 3.3. In stage 1 of the abstraction path, $Y_{out'}$ is divided into non-overlapping patches of dimension $\hat{H}\times\hat{W}\times\hat{D}$ each. These are further flattened into 1D vectors for projection onto a different embedding space, with dimension $\hat{C}$, using a linear embedding layer. Here, $\hat{C}$ was selected as 48. The transformed vectors are subsequently subjected to Layer Normalization before being provided as input to θ. The vectors are organized into several local windows in θ, with the inter-dependencies between them being calculated using a self-attention component α_{sa} within each local window as

$$\alpha_{sa}(Q,K,V) = \eta(\frac{QK^T}{\sqrt{dim}})\cdot V, \quad (3.6)$$

where Q, K, V, denote the query, key, and value vectors, respectively, dim

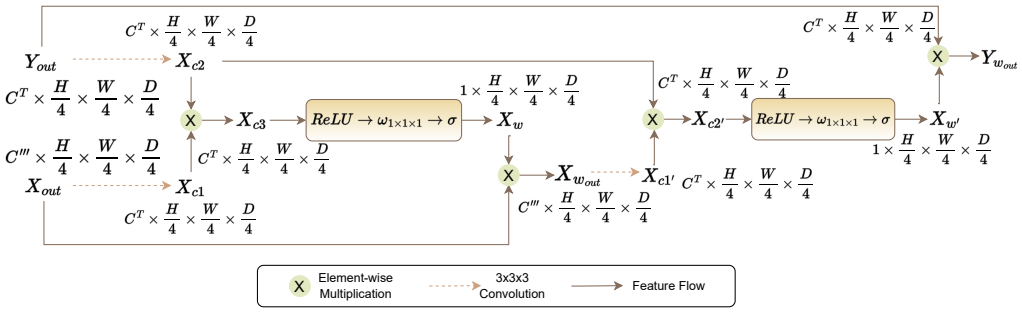

FIGURE 3.4: Schematic view of *Cross-Context Attention*.

corresponds to the size of these vectors, and η is the softmax activation function. Window size $7 \times 7 \times 7$ was used in experiments. Output of θ is passed through Layer Normalization, followed by processing by an MLP layer (Fig. 3.3).

For the second component θ', the windows are shifted using 3D cyclic shifting (as in Ref. [76]) to model interdependencies among patches between windows. Neighboring patches are grouped together by stacking them along the channel dimension – in order to reduce the spatial dimension of the input volume by a factor of 2 to produce the output volume $Y_{out} \in C^T \times \frac{H}{4} \times \frac{W}{4} \times \frac{D}{4}$.

The *DA* module in subsequent levels 2-4 of the abstraction path involves amplification of relevant feature maps [eqns. (3.3)-(3.5)], along with the θ and θ' components to hierarchically capture global contextual information.

3.2.3 Cross-Context Attention

The *DA* module generates feature maps incorporating spatial information, while the *SpectraConv* module provides a feature volume containing relevant wavelet information from the same image. Due to discrepancies between domains, *WaveCoformer* must bridge the gap to learn the full spectrum representation of global and local features. The *Cross-Context Attention (CCA)* module combines spectral and spatial representations to fully represent target anatomical structures. Fig. 3.4 illustrates the *CCA* module.

Initially, X_{out} and Y_{out} are simultaneously subjected to a convolution $w_{3\times3\times3}$. This step refines, aligns and transforms the features to enhance interaction and integration within the *CCA* module. The additional non-linearity introduced by convolution helps to learn more meaningful and complex features from both domains for better integration. Furthermore, it modifies the total number of channels, improving alignment and interaction between the two feature volumes.

The output feature volumes, $X_{c1} \in \mathbb{R}^{C^T \times \frac{H}{4} \times \frac{W}{4} \times \frac{D}{4}}$ and $X_{c2} \in \mathbb{R}^{C^T \times \frac{H}{4} \times \frac{W}{4} \times \frac{D}{4}}$, are multiplied to get $X_{c3} \in \mathbb{R}^{C^T \times \frac{H}{4} \times \frac{W}{4} \times \frac{D}{4}}$. The element-wise multiplication results in the amplification of locations that exhibit higher response values in both volumes. Conversely, positions exhibiting lower response values in both volumes get attenuated; thereby, eliminating the influence of locations characterized by weaker or noisy responses. It helps capture interactions between spatial and frequency information. Next, X_{c3} undergoes the activation of *ReLU* to mitigate any negative value(s). This is followed by a convolution $w_{1\times1\times1}$ and a sigmoid activation σ to generate the final weight map $X_w \in \mathbb{R}^{1 \times \frac{H}{4} \times \frac{W}{4} \times \frac{D}{4}}$. The

weight map assigns weights to the features of the wavelet domain based on their significance in the spatial context. The calibrated output $X_{w_{out}} \in \mathbb{R}^{C''' \times \frac{H}{4} \times \frac{W}{4} \times \frac{D}{4}}$ is represented as

$$X_{w_{out}} = X_w \otimes X_{out}. \tag{3.7}$$

$X_{w_{out}}$ is subsequently processed, according to the previous steps, to generate the score map $X'_w \in \mathbb{R}^{1 \times \frac{H}{4} \times \frac{W}{4} \times \frac{D}{4}}$. This score map finally enhances Y_{out}; thereby, resulting in a comprehensive representation that incorporates both spatial and frequency information. The final output of the *CCA* module $Y_{w_{out}} \in \mathbb{R}^{C^T \times \frac{H}{4} \times \frac{W}{4} \times \frac{D}{4}}$ is generated as

$$Y_{w_{out}} = \sigma[w_{1\times1\times1}\{ReLU(w_{3\times3\times3}\overline{X_{w_{out}}} \otimes X_{c2})\}] \otimes Y_{out}. \tag{3.8}$$

Here, the dynamic weighting strategy enables the adjustment of the contribution of both spatial and spectral domains, depending on the specific characteristics of the input image and the target region. This enables the network to adjust to various contexts to acquire more refined feature representations.

3.3 Implementation Details

This section presents the datasets and preprocessing steps used to evaluate the generalizability of *WaveCoformer*, along with the loss functions for training.

3.3.1 Datasets

WaveCoformer was trained and validated using the *Synapse* and *AdrenalSeg* datasets. The details of the individual data are provided in Sec. 1.6.1. The preprocessing steps for the *Synapse* and *ACDC* datasets are the same as discussed in Sec. 2.3.2.

3.3.2 Loss function

A hybrid loss function consisting of Dice loss ($\mathbf{L_{dice}}$) and Categorical Cross-Entropy loss ($\mathbf{L_{cce}}$) [124] was used to train *WaveCoformer*.

The Dice loss function is frequently used in image segmentation tasks involving class imbalance. This scenario is common in medical image segmentation, where the number of pixels related to the target structure to be demarcated is significantly lower than the number of background pixels. Mathematically $\mathbf{L_{dice}}$ is expressed as

$$\mathbf{L_{dice}} = \kappa - \sum_{\kappa=1}^{K} \left(\frac{2 \sum_{i=1}^{V} \hat{y}_{\kappa,i} y_{\kappa,i} + \epsilon}{\sum_{i=1}^{V} \hat{y}_{\kappa,i} + y_{\kappa,i} + \epsilon} \right), \tag{3.9}$$

where c denotes the total number of classes, with $\hat{y}_{\kappa,i}$, $y_{\kappa,i}$ being the predicted and ground truth values (respectively) for the ith voxel w.r.t class κ, V is the total number of voxels in the input, ϵ is the additive smoothing parameter introduced to prevent any division-by-zero error, and DSC [1.21] is the Dice-score coefficient. K denotes the total number of classes.

TABLE 3.1: Comparative performance of *WaveCoformer* on *Synapse* dataset, while varying weightage of *DA* and *SpectraConv* modules.

Weightage factors for SpectraConv (x) and DA (y)	**Mean *DSC***
x = 0.2 y = 0.8	0.8207
x = 0.4 y = 0.6	0.8279
x = 0.6 y = 0.4	0.8338
x = 0.8 y = 0.2	0.8158
x = 0.5 y = 0.5	**0.8404**

The categorical cross-entropy loss is a variant of the cross-entropy loss that is specifically designed for multiclass segmentation. It quantifies the difference between the probability distributions of the predicted segmentation map and the ground truth. It is represented as

$$\mathbf{L}_{cce} = -\frac{1}{V}\sum_{i=1}^{V}\sum_{\kappa=1}^{c} y_{\kappa,i}\log\left(\hat{y}_{\kappa,i}\right). \tag{3.10}$$

The composite loss function, incorporating both Dice and Categorical Cross Entropy loss, is advantageous as it allows leveraging the respective benefits of each component. While $\mathbf{L_{dice}}$ effectively addresses the issue of class imbalance between foreground and background pixels, the $\mathbf{L}_{cce}$ component compensates the trade-off between False Positives (FP) and False Negatives (FN) on the predicted output map [124]. The composite loss function is expressed as

$$\mathbf{L}(\{\hat{\mathbf{\Gamma}}, \mathbf{\Gamma}\}; \mathbf{\Omega}) = \mathbf{L_{dice}}(\{\hat{\mathbf{\Gamma}}, \mathbf{\Gamma}\}, \mathbf{\Omega}) + \mathbf{L_{cce}}(\{\hat{\mathbf{\Gamma}}, \mathbf{\Gamma}\}, \mathbf{\Omega}), \tag{3.11}$$

where $\hat{\Gamma}$ and Γ represent the predicted segmentation map and the ground truth, respectively, and Ω corresponds to the model parameters.

3.4 Results and Discussion

Table 3.1 displays the average values of *DSC* obtained across different abdominal organs by adjusting the weights assigned to the output volumes of the modules *DA* and *SpectraConv*. This experiment examines the importance of the spectral and spatial components in the segmentation of different organs. Optimal performance was achieved by assigning equal weights to both modules. This implies that an equal contribution from both modules is essential for precise segmentation across different organs.

A series of experiments was conducted to analyze the role of various components, in achieving an effective performance of *WaveCoformer*, and is depicted in Table 3.2. *WaveCoformer* with all the modules (*SpectraConv, DA* and *CCA*) obtained a mean *DSC* of 84.04% for all abdominal organs. This illustrates the efficacy of the integrated methodology in precisely segmenting target structures.

TABLE 3.2: Comparative study of performance metrics, over different variants of *WaveCoformer*, on *Synapse* dataset.

Model variants	DSC										
	Spleen	Right kidney	Left kidney	Gall Bladder	Liver	Pancreas	Stomach	Right Adrenal	Left Adrenal	Mean	Wilcoxon Test (P-value)
Proposed architecture	**0.9536**	**0.9421**	**0.9398**	0.7045	**0.9666**	**0.7888**	**0.9073**	**0.6868**	**0.6742**	**0.8404**	-
No SpectraConv	0.9265	0.9246	0.9267	0.7156	0.9572	0.6465	0.7513	0.6738	0.5746	0.7885	<0.05
No Dual Attention	0.8958	0.9042	0.8941	0.6472	0.9474	0.6399	0.7303	0.618	0.5952	0.7636	<0.05
No Cross Context Attention	0.9033	0.9223	0.9193	**0.7524**	0.9521	0.7414	0.7551	0.6223	0.5774	0.794	<0.05
No Channel Attention	0.897	0.9194	0.9132	0.6786	0.9536	0.6358	0.7418	0.6462	0.5243	0.7700	<0.05
No Swin Transformer	0.9048	0.9108	0.9181	0.7204	0.955	0.7207	0.7739	0.6351	0.5439	0.7900	<0.05

Removal of the module *SpectraConv* significantly decreased the mean Dice score from 84.04% to 78.85%. The reduction of 5.2% highlights the significant influence of the spectral attributes (or the wavelet coefficients) in achieving accurate segmentation. Performance declines (ranging from 1.31% - 9.96%) were observed in smaller organs such as the kidneys, adrenal glands, and pancreas. This underscores the importance of *SpectraConv* in identifying the subtle characteristics of these challenging-to-segment structures. A notable decrease in *DSC* of the larger organs also suggests the role of the proposed module in the capture of high-frequency details (related to textural information) for better segmentation quality.

Elimination of *DA* resulted in the most pronounced decline in performance in all organs. The mean Dice score shows a reduction of 7.7%, suggesting that the incorporation of channel attention and global spatial dependencies is crucial to achieve precise segmentation. Removal of *DA* has a greater effect on smaller organs such as the gall bladder (Dice score decrease of 5%) and adrenal glands (7%–8%). This suggests that the ability of *DA* to focus and capture contextual information is essential for accurately segmenting these smaller and more difficult structures. The decline in performance in larger organs, such as the liver, indicates that *DA* has a significant role in maintaining organ continuity and capture of global shape characteristics.

Although inclusion of both components substantially improved the overall performance of *DA* within the network, the ablation study uncovered a notable interaction between them. Eliminating the Swin transformer, while retaining solely the channel attention component, resulted in an improvement of approximately 3% (relative to the variant lacking *DA* module). In contrast, eliminating channel attention while retaining only the Swin Transformer led to an enhancement of approximately 1% relative to the variant devoid of the *DA* module. Hence, it may be inferred that the incorporation of channel attention significantly enhances the overall performance in *WaveCoformer*. The targeted emphasis on pertinent feature maps results in an enhanced feature representation for effective segmentation.

The fourth row of Table 3.2 illustrates the performance decrease, after the removal of the *CCA* module. The overall *DSC* decreased by 4.6%, highlighting the importance of *CCA* in improving the overall performance of the model. The performance decline for individual organs varied between 1% and 15%. Therefore, the module *CCA* is found to be essential for the synergistic integration of the global context, encoded in spatial features, along with fine textural details embedded in spectral features. Combining both types of features, an integrated

TABLE 3.3: Summary of the impact of various components of *WaveCoformer* on segmentation of *Synapse* dataset.

Attention Module	Description	Effect on Segmentation Results
SpectraConv	Extracts relevant features from wavelet coefficients	- Improves the overall DSC by 5.2 % (when included) - Significantly improves the segmentation of smaller organs with complex textures or fine details like kidneys (1-2 % increase), adrenal glands (1-10% increase) and pancreas (14% increase).
DA (full module)	Filters relevant feature maps, followed by capture of global spatial dependencies	- Improves the overall DSC by 7.7 % (when included) - Essential for segmenting both small and large-sized organs. Significantly improves the performance over small organs like kidneys (4% increase), adrenal glands (8 - 9% increase), pancreas (14% increase), and large organs, *viz.* liver (2% increase) - Effectively captures contextual information for accurate segmentation of smaller organs. - Maintains organ continuity and captures global shape characteristics.
CCA	Fuses spatial and spectral features.	-Improves overall accuracy by combining complementary information. - Essential for combining both feature types to effectively capture the overall structure of large organs along with the subtler details of smaller-sized organs. Removal of the module shows significant decline in performance across all organs.

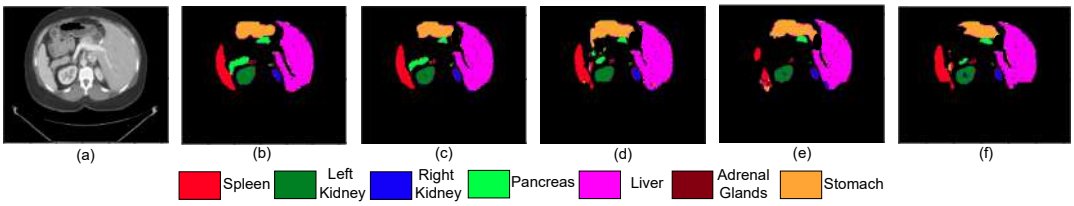

FIGURE 3.5: Segmentation maps from different variants of *WaveCoformer*. (a) Input, (b) Ground Truth, (c) *WaveCoformer* with full configuration, (d) *WaveCoformer* w/o *SpectraConv*, (e) *WaveCoformer* w/o *DA* and (f) *WaveCoformer* w/o *CCA*.

approach is essential to effectively capture the overall structure of large organs with subtler details of smaller organs. The last column of Table 3.2 reports the results of the Wilcoxon signed rank test to compare the results of *DSC* of the full configuration of the model versus the different variants listed. The $p - value < 0.05$ in all cases indicates the importance of every module in improving overall performance. Table 3.3 summarizes the impact of the different modules on the overall performance of *WaveCoformer*.

Fig. 3.5 presents the sample segmentation outputs of the aforementioned variants of *WaveCoformer*. A significant number of *FN* is observed in the segmented area associated with the pancreas when *SpectraConv* is absent in Fig. 3.5(d). The abdominal area exhibits many *FP*s. In Fig. 3.5(e), multiple discrepancies are observed in both small and large organs such as the spleen, pancreas, and kidneys that are consistent with the quantitative findings regarding the effects of removing *DA* presented in row 3 of Table 3.2. On removing *CCA*, several mis-segments are observed in the corresponding output Fig. 3.5(f).

Performance of *WaveCoformer* was compared (over two publicly available datasets) with state-of-the-art algorithms in the literature, in terms of metrics *DSC* [eq. (1.21)], *IoU* [eq. (1.20)], *HD*95 [eq. (1.22)]. The results are summarized in Table 3.4 and Fig. 3.6. *WaveCoformer* was observed to generate the highest *DSC* values during segmentation of larger organs (such as the spleen, liver, and stomach), as well as some smaller organs (such as the kidneys, pancreas, and left adrenal gland), as shown in Table 3.4. This indicates our model can generalize well to anatomical structures exhibiting varying shapes and sizes. It is robust in delineating target regions that differ in structure among patients, as is evident from the lowest variability of the metric scores in Fig. 3.6(a) across

TABLE 3.4: Comparison of *WaveCoformer* with baseline models, on *Synapse* and *AdrenalSeg*.

Model	*Synapse*												*AdrenalSeg*		
	DSC										Mean IoU	Mean HD95	DSC	IoU	HD95
	Spleen	Right kidney	Left kidney	Gall Bladder	Liver	Pancreas	Stomach	Right Adrenal	Left Adrenal	Mean					
U-Net	0.91	0.90	0.92	0.56	0.96	0.70	0.78	0.61	0.59	0.77	0.67	38.83	0.68	0.71	60.22
V-Net	0.89	0.92	0.92	0.59	0.95	0.75	0.80	0.60	0.45	0.76	0.65	25.29	0.68	0.69	56.49
U-Net++	0.92	0.91	0.89	0.69	0.95	0.75	0.79	0.52	0.10	0.73	0.62	55.27	0.63	0.50	72.59
Attention *U*-Net	0.91	0.88	0.88	0.58	0.96	0.55	0.79	0.60	0.54	0.74	0.63	51.61	0.70	0.57	81.82
U−Net with EfficientNet-b0	0.85	0.90	0.88	0.64	0.90	0.50	0.71	0.60	0.52	0.72	0.60	67.47	0.65	0.52	56.38
U−Net with EfficientNet-b1	0.74	0.80	0.83	0.55	0.90	0.54	0.61	0.52	0.50	0.67	0.55	75.85	0.61	0.52	95.61
U−Net 3+	0.87	0.92	0.86	0.67	0.94	0.69	0.73	0.57	0.48	0.74	0.62	54.43	0.69	0.56	48.01
Swin UNETR	0.95	0.93	0.92	**0.76**	0.96	0.80	0.80	0.69	0.62	0.83	0.73	13.99	0.71	0.58	51.19
UNETR	0.89	0.90	0.89	0.53	0.95	0.67	0.79	0.53	0.50	0.74	0.63	26.94	0.60	0.48	100.95
TransUNET	0.85	0.89	0.79	0.76	0.59	0.72	0.88	0.59	0.47	0.72	0.67	36.81	0.73	0.60	34.79
TransAttUNet	0.90	0.87	0.91	0.49	0.94	0.59	0.66	0.65	0.59	0.73	0.65	15.81	0.67	0.56	108.00
DS-TransUNET	0.49	0.61	0.52	0.53	0.67	0.52	0.58	0.57	0.53	0.56	0.57	27.58	0.56	0.45	**19.70**
Swin UMamba	0.87	0.90	0.91	0.42	0.91	0.62	0.66	0.59	0.56	0.72	0.63	20.1	0.64	0.55	71.13
WaveCoformer	**0.95**	**0.94**	**0.94**	0.70	**0.97**	**0.80**	**0.90**	**0.69**	**0.67**	**0.84**	**0.73**	**11.16**	**0.83**	**0.73**	42.70

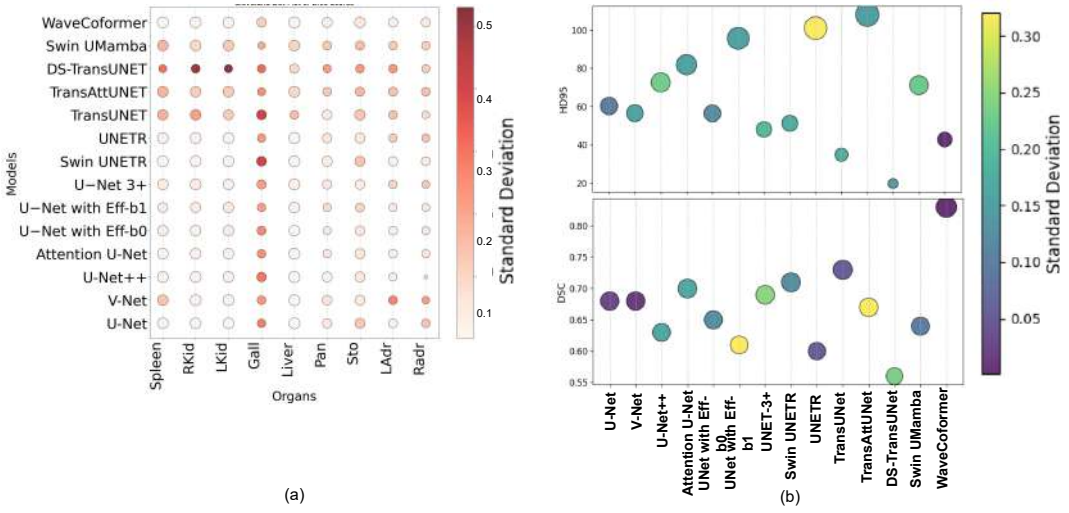

FIGURE 3.6: Visualization of (a) *DSC* on *Synapse* and (b) *DSC* and *HD*95 on *AdrenalSeg* datasets. Radius of the circle denotes the metric values, while the intensity of the color signifies the standard deviation. Colour intensity scale indicates the corresponding standard deviation value.

all samples. *WaveCoformer* also produced the lowest mean *HD*95 and *IoU*, indicating a close alignment of the predicted with the actual boundaries of the target regions.

Fig. 3.7 serves to validate these claims, based on sample segmentation maps displaying the different abdominal organs. *WaveCoformer* exhibited a significantly higher degree of similarity to ground truth compared to the maps generated by other baseline models, as visualized in the figure. *FP* regions corresponding to the stomach are present in the Swin UNETR and UNETR results, as shown in the first row of Figs. 3.7 (d) and (e). The results obtained from the *V*-Net in the top row of Fig. 3.7 (f) indicate that the pixels corresponding to the boundary regions of the spleen were misclassified. In addition, it has an inability to accurately identify the anatomical regions associated with the left adrenal and right adrenal glands. The results in the second row of Figs. 3.7 (d) and (f) show an under-segmented right kidney.

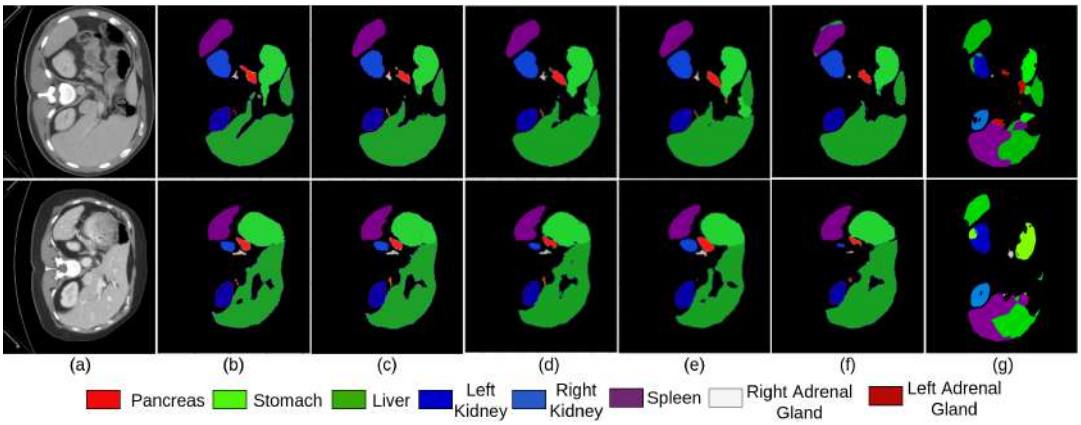

FIGURE 3.7: Sample segmentation maps comparing the performance of *WaveCoformer*, with other baseline architectures, on *Synapse* dataset. (a) The CT slice, with corresponding (b) Ground truth, and output from (c) *WaveCoformer* (d) Swin UNETR (e) UNETR (f) *V*-Net and (g) TransAttUNet.

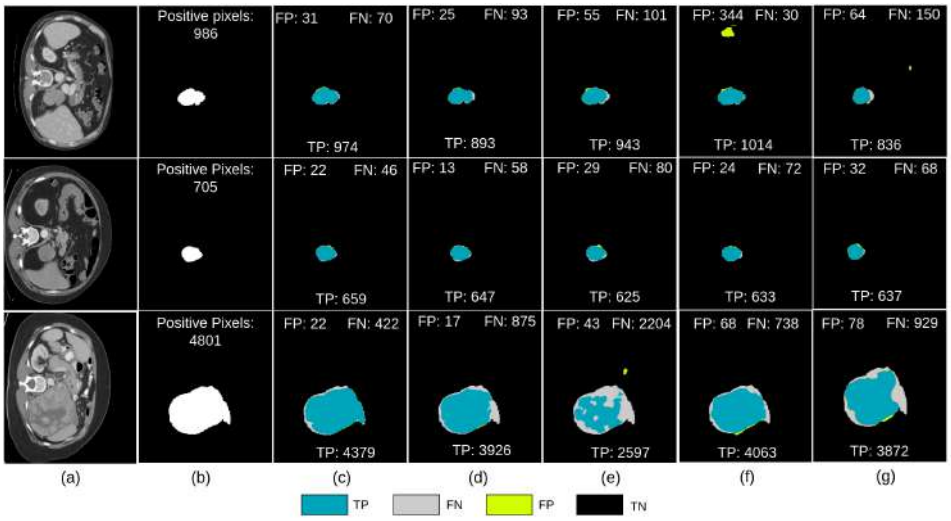

FIGURE 3.8: Sample segmentation maps comparing the performance of *WaveCoformer*, with other baseline architectures, on *AdrenalSeg* dataset. (a) Input CT image (b) Ground truth of the tumor, and corresponding output from (c) *WaveCoformer* (d) Swin UNETR (e) UNETR, (f) *V*-Net and (g) TransAttUNet.

For the gall bladder, *WaveCoformer* yielded the second-best result in terms of *DSC*. This may be due to a variable and poorly defined shape, as well as interference from adjacent tissues and organs, such as the liver and intestines, which exhibited similar intensity levels. This is evident from the larger standard deviation value w.r.t *DSC*, as shown in Fig. 3.6(a). Although *WaveCoformer* has the second highest value *DSC*, the lowest standard deviation compared to all baselines shows that the proposed approach is robust to variations in relation to the morphological structure of the gall bladder with its combination of spatial and frequency information.

WaveCoformer achieved the highest scores of 83% and 73% for *DSC* and

TABLE 3.5: Comparing *WaveCoformer* with other baseline models, in terms of TFLOPs and number of parameters (in millions)

Model	Parameters	FLOPs
U-Net	90.294 M	3.587 T
V-Net	45.646 M	1.453 T
U-Net++	97.626 M	2.26 T
Attention *U*-Net	90.645 M	3.616 T
U-Net with EfficientNet-b0 backbone	60.566 M	2.024 T
U-Net with EfficientNet-b1 backbone	63.342 M	2.026 T
U-Net 3+	80.192 M	25.988 T
Swin UNETR	61.990 M	1.319 T
UNETR	115.184 M	2.259 T
TransUNET	109.54 M	0.566 T
DS-TransUNET	171.339 M	0.377 T
TransAttUNet	30 M	0.71 T
Swin UMamba	55.05 M	2.102T
WaveCoformer	**25.852 M**	**0.184 T**

IoU, surpassing the other approaches listed. This underscores its ability to accurately delineate the tumor position, even amongst considerable size discrepancies among various patients, as visualized in Fig. 3.8. The sample outputs show that the *WaveCoformer* has a higher TP pixel count compared to other models. Although in the first row of Fig. 3.8 (e), it can be seen that *V*-Net has a comparatively higher TP pixel count than *WaveCoformer*, this model also captures several FP regions. This indicates that *V*-Net model cannot effectively differentiate between the tumor and other abdominal structures present in the input. Swin UNETR and UNETR produce undersegmented tumor regions, as can be observed in the third row of Fig. 3.8 (c) and (d).

However, *WaveCoformer* produced the third-best result w.r.t. the $HD95$ scores in Table 3.4. Although *WaveCoformer* effectively represented the total tumor volume, there may be specific segmented boundaries that diverge from the actual data. This significant variability in the size and morphology of adrenal tumors, within the *AdrenalSeg* dataset, poses a challenge to its generalizability. This might result in boundary discrepancies in certain instances. Although *WaveCoformer* achieved the third-best performance, the low standard deviation scores demonstrate its robustness, as evident in Fig. 3.6 (b). The model with the lowest $HD95$ score had a comparatively larger standard deviation with respect to $HD95$ compared to *WaveCoformer*.

Table 3.5 provides an analysis of the parameter count and Tera FLoating point OPerations (TFLOPs) of *WaveCoformer* with the other baseline architectures considered. The study reveals that *WaveCoformer* exhibits the lowest parameter count and TFLOP, compared to these algorithms. This validates the claim regarding the higher computational efficiency of our *WaveCoformer*. It produced a reduced carbon footprint for ecologically sustainable solutions [31].

Fig. 3.9(a) illustrates the various weights generated by the channel amplification block in the *DA* module. This helps identify pertinent maps from the feature map volume. The columns associated with channel indices, in shades of red and yellow, indicate channels that convey highly relevant information on the target structures. The black columns correspond to channels of minimal significance. Fig. 3.9(b) illustrates the attention matrix derived from the Swin Transformer block of the *DA* module, highlighting the relationships among patches within a window size of 7×7. The colors represent the strength of

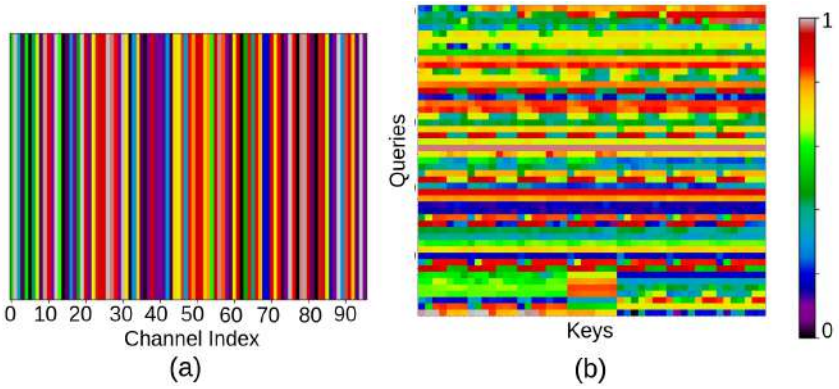

FIGURE 3.9: Visualization of (a) channel weights from the salient map amplification sub-module and (b) attention map from the Swin transformer block of *DA* module. The colorbar represents the weight values $\in [0, 1]$.

attention, with red shades indicating strong dependence between the $Q - K$ pair, while dark blue and black shades denote weaker dependencies. The wide distribution of attention weights indicates that the block effectively captures a broader spectrum of global dependencies. The horizontal bands present in the matrix indicate that a single query is engaging with multiple adjacent keys. This suggests that a query considers information from a broader context.

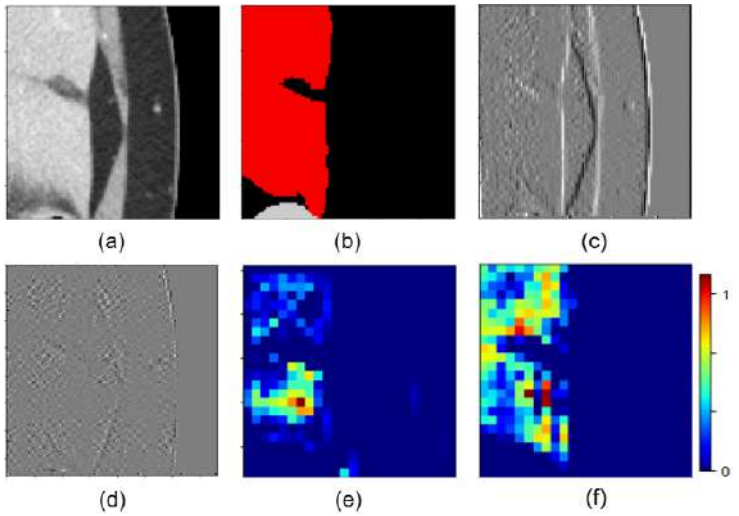

FIGURE 3.10: Feature maps from *SpectraConv* and *CCA* module. Sample (a) Input image patch with (b) corresponding ground truth, (c) and (d) high-frequency wavelet coefficients, feature maps from (e) *SpectraConv* module, (f) *CCA* module.

Fig. 3.10 shows sample activation maps of the modules *CCA* and *SpectraConv*, which help to interpret their internal mechanisms. Fig. 3.10(f) demonstrates stronger and denser activation in spatial areas that align with the ground truth map shown in Fig. 3.10(b). This highlights the effectiveness of *CCA* in producing strong representations of target structures. The denser activations reflect the ability of the model to represent the target regions with a high level

of detail, effectively capturing both intricate details and overall structures. In Fig. 3.10(e), the activations correspond to the learning of information related to the high-frequency components of the target region, as seen in the wavelet coefficient map in Fig. 3.10(c) and (d). Distinct activations are visible along the boundary regions. This corresponds to the representation of relevant information related to the edges and textures by the *SpectraConv* module from the wavelet coefficients.

The enhanced performance of *WaveCoFormer* results from the synergistic integration of its essential components, *viz.* modules *DA*, *SpectraConv*, and *CCA*. The DA module, which incorporates channel attention and global context capture, facilitates the generation of a refined representation of organs characterized by diverse shapes and sizes. The *SpectraConv* effectively captures nuanced textural details that are critical for smaller structures. The module *CCA* integrates features from both the spatial and wavelet domains to create a robust image representation. These modules capture and integrate multiscale information from both spatial and frequency domains, enabling the model to achieve accurate and robust segmentation across different organs and datasets. This highlights the effectiveness of *WaveCoFormer* as a tool for medical image segmentation and its potential to improve clinical practice.

3.5 Conclusion

This chapter presented the development and validation of a novel wavelet-infused hybrid deep learning framework *WaveCoformer*. The model simultaneously analyzed features from both the spatial and spectral domains of volumetric medical images. The *SpectraConv* module acquired multi-scalar textural patterns and granular details from the DWT of the input images. The *DA* block identified the relevant activation maps from the feature maps obtained from the spatial domain of the image. This was followed by determining the global dependencies among the different spatial locations, using the self-attention mechanism of the transformer module. The *CCA* block integrated the complementary features of the convolution and transformer blocks.

The integration of information from both the frequency and spatial domains resulted in a superior segmentation performance. The hybridization of convolution and transformer mechanisms generated a complete and information-rich representation through a combination of intricate textures with coarser global structural information. This enhanced the capabilities of our *WaveCoformer*. The model also demonstrated the lowest TFLOP value and the total parameter count, as observed from the comparative study with related baseline architectures. This was due to the focus achieved by the *DA* module in identifying relevant activation maps, for subsequent processing by the transformer block. The minimized storage requirements make it tenable for deployment in diverse computing environments. However, the hybrid design leads to a higher practical inference latency compared to the fully convolutional models.

The dependence of *WaveCoformer* on the self-attention mechanism suffers from the quadratic scaling of the computational complexity, with respect to the

input size. This forms a bottleneck for processing high-resolution volumetric data. The following chapter explores an alternative approach, by leveraging the linear complexity and constant memory usage of Vision-xLSTM to model global context for efficient and precise medical image segmentation.

Chapter 4

Optimizing Segmentation with Vision Extended LSTM

"Simplicity is the ultimate sophistication."

— Leonardo da Vinci

4.1 Introduction

Recent medical image segmentation frameworks leverage the strengths of both CNNs and ViTs to achieve effective performance. CNNs excel at capturing fine-grained details, whereas ViTs efficiently model the long-range dependencies between different regions of the input image. The previous chapter discussed the efficiency of a hybrid CNN-Transformer model which utilized both spatial and spectral domain characteristics for high-quality segmentation. However, the central mechanism of Transformers presents a major implementation bottleneck. The self-attention mechanism, responsible for modeling global dependencies, suffers from quadratic computational complexity $\mathcal{O}(N^2)$ with respect to the input sequence length N. This leads to a substantial requirement for GPU processing and memory while processing high-dimensional volumetric medical images, *viz.* CT and MRI scans.

In order to address this gap, Chapter 4 explores Vision Extended Long-Short Term Memory (ViL) [3], a variant of xLSTM for vision-related tasks, as a promising alternative to ViT. The ViL has a linear computational complexity $\mathcal{O}(N)$ and constant memory complexity $\mathcal{O}(1)$, with respect to N. This makes ViL a strong candidate for a resource-efficient powerful segmentation backbone, capable of learning long-range dependencies without the computational burden associated with ViTs.

This chapter details the two-stage investigation of hybridizing CNN with ViL towards developing a robust and efficient segmentation model. The *U-Vision-xLSTM (U-VixLSTM)* [35] is the initial model developed by embedding ViL within the $U-$shaped segmentation framework. The ViL blocks capture the temporal and global relationships, within the fine-grained details extracted from the CNN feature maps. An advanced framework, called *Rotational U-Vision-xLSTM (Rot-UViL)* [38], is thus developed to capture cross-dimensional dependencies in a computationally efficient manner within the CNN activation volumes. This enables the ViL to model global correlations across channel, width and height dimensions for a comprehensive contextual understanding of the target anatomical structures in a volumetric input. The contribution of the chapter is summarized below.

- Development of *U-VixLSTM* by integrating CNNs with ViLs. CNNs capture fine-grained textural information and local patterns, corresponding to the target anatomical structures, from the input image. The ViL block encodes the global context within the intermediate output volumes, as obtained from the CNN layers.

- Advancing the above framework to *Rot-UViL*, which captures the cross-dimensional dependencies for a holistic understanding of global context within the volumetric feature maps.

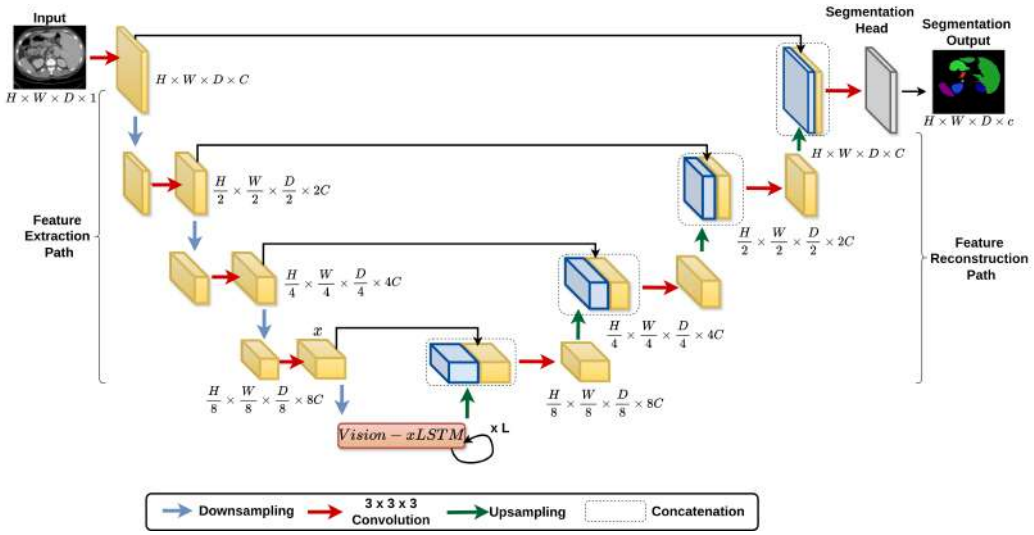

FIGURE 4.1: Architectural framework of *U-VixLSTM*.

- Experimental results on multiple publicly available datasets illustrate the effectiveness of the developed models, both in terms of performance and utilization of computing resources.

The subsequent part of this chapter is organized as follows. Section 4.2 describes the methodology of *U-VixLSTM*, with the implementation details and experimental findings to validate its effectiveness. Section 4.3 details the advancement of this framework to a sophisticated version *Rot-UViL*, along with a comprehensive experimental evaluation on volumetric medical images. Section 4.4 presents a detailed comparative analysis among the different segmentation models described so far. Finally, Section 4.5 concludes the chapter.

4.2 U-Vision-xLSTM

The structural framework of *U-VixLSTM* model is depicted in Fig. 4.1. It follows the classic *U*-shaped framework, characterized by feature extraction and reconstruction pathways. Although the approach is outlined within the framework of 3D volumetric image processing, it can easily be modified for 2D images by reducing spatial dimensions.

The feature extraction arm has multiple layers of CNNs with ViL blocks in the bottleneck. Each ViL block contains the mLSTM layer [11] to capture long-range dependencies along with temporal awareness. The mLSTM layer employs an exponential gating mechanism to strike a balance between retaining past information and integrating new inputs. The ViL block processes the feature volumes from CNNs to generate an abstract high-level representation of the image. The reconstruction path gradually builds the high-dimensional segmentation output using the contextual representation from the ViL block. The output of each convolution block in the feature extraction path is directed to its corresponding counterpart, at the same level of the feature reconstruction path, through skip connections. This facilitates the integration of feature maps,

that originate at various levels of abstraction, with the activation maps at the corresponding level of the decoder. It ensures a judicious combination of finer textural details from earlier convolution levels with coarser semantic information of deeper levels; thereby, resulting in enhanced context-sensitive predictions.

4.2.1 Feature extraction

This path consists of key components, *viz.* CNN for high-level feature learning and ViL for capturing global dependencies. The volumetric input image $I \in \mathbb{R}^{H \times W \times D \times 1}$ is passed through a series of convolution layers to hierarchically construct an intermediate abstract and high-level representation of the image denoted by $x \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times \frac{D}{8} \times 8C}$. Here, H, W and D represent the height, width, and depth of the intermediate feature volume, respectively, with C corresponding to the number of channels.

Next, x is divided into $P \times P \times P$ non-overlapping patches. This is followed by the flattening of these patches into 1D vectors to yield a tokenized representation $\mathbf{t} \in \mathbb{R}^{N \times (P^3 \frac{C}{4})}$. Here, $N = (\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4})/P^3$ denotes the number of flattened patches with dimension $P^3 \frac{C}{4}$. The flattened patches $(t^1, t^2, \ldots, t^N)$ are then projected into a Z-dimensional embedding space, with learnable positional embeddings being added to preserve spatial location information. Mathematically, this is expressed as

$$\mathbf{p} = [t^1\mathbf{K}; t^2\mathbf{K}; \ldots; t^N\mathbf{K}] + \mathbf{K_{pos}}, \tag{4.1}$$

where $\mathbf{K} \in \mathbb{R}^{(P^3 \frac{C}{4} \times Z)}$ is the projection matrix and $\mathbf{K_{pos}} \in \mathbb{R}^{N \times Z}$ is the position embedding matrix, with $\mathbf{p} \in \mathbb{R}^{N \times Z}$ representing the matrix of flattened patches.

The projected patches are processed by the ViL blocks, with the even-numbered blocks handling patch tokens from the top left to the bottom right, and the odd-numbered ones from the bottom right to the top left. Such bidirectional processing enables the ViL to capture robust global dependencies in the input. The detailed processing steps of ViL were discussed in Sec. 1.3.2.

Lemma 2 (Computational efficiency of ViL)**.** *Let $\mathbf{p} \in \mathbb{R}^{N \times Z}$ be an input token sequence, with N the number of patches and Z the embedding dimension. For vision tasks with $N >> Z$, ViL is asymptotically more efficient than ViT by a factor of $\mathcal{O}(\frac{N}{Z})$.*

Proof. The self-attention mechanism is defined as

$$\alpha_{sa} = \eta(\frac{QK^T}{\sqrt{Z}}).V,$$

where $Q \in \mathbb{R}^{N \times Z}$, $K \in \mathbb{R}^{N \times Z}$ and $V \in \mathbb{R}^{N \times Z}$ are the query, key and value matrices respectively. The matrix multiplication QK^T results in an attention score matrix of dimension $\mathbb{R}^{N \times N}$. The associated computational cost $= \mathcal{O}(N^2 Z)$. Subsequently, the attention score is multiplied with V to get the final attention score matrix $\alpha_{sa} \in \mathbb{R}^{N \times Z}$. Therefore, the final computational complexity of the self-attention mechanism becomes $= \mathcal{O}(N^2.Z) + \mathcal{O}(N^2.Z) \equiv \mathcal{O}(N^2.Z)$.

The ViL block processes one token $x_t \in \mathbb{R}^{N \times Z}$ at each timestep t. The major operations involve multiplication with weight matrices $\mathbf{W_{Q_t}} \in \mathbb{R}^{Z \times Z}$, $\mathbf{W_{K_t}} \in \mathbb{R}^{Z \times Z}$, $\mathbf{W_{V_t}} \in \mathbb{R}^{Z \times Z}$, $\mathbf{W^I} \in \mathbb{R}^{Z \times Z}$, $\mathbf{W^F} \in \mathbb{R}^{Z \times Z}$ to generate query, key and value matrices with input and forget gate activations.
$\therefore$ the total computational cost for processing $x_t = \mathcal{O}(Z^2) + \mathcal{O}(Z^2) + \mathcal{O}(Z^2) + \mathcal{O}(Z^2) + \mathcal{O}(Z^2) \equiv \mathcal{O}(Z^2)$;
$\forall$ N number of patches the cost becomes $\mathcal{O}(NZ^2)$.
$\therefore$ Gain ratio $= \frac{\text{Cost of } \alpha_{sa}}{\text{Cost of } ViL} = \frac{\mathcal{O}(N^2.Z)}{\mathcal{O}(NZ^2)} = \mathcal{O}(\frac{N}{Z})$. □

4.2.2 Feature reconstruction

A trilinear upsampling operation τ is employed, at every level l, to increase the spatial dimension of the feature maps obtained from the previous level $l+1$. This helps align the spatial dimensions of the feature maps with those received from the corresponding level l of the feature extraction path. The upsampled feature map at level l is expressed as

$$\mathbf{U}_l = \tau(\mathbf{F}_{l+1}), \tag{4.2}$$

where $\mathbf{F}_{l+1}$ is the output volume from level $l+1$. The upsampled feature map $\mathbf{U}_l$ is then concatenated with the feature volume from the corresponding level l of the feature extraction path, denoted as $\mathbf{E}_l$. The concatenated feature volume at level l is given by

$$\mathbf{C}_l = \text{Concat}(\mathbf{U}_l, \mathbf{E}_l). \tag{4.3}$$

It is then convolved to yield the output volume $\mathbf{R}_l$ at level l of the feature reconstruction path, as

$$\mathbf{R}_l = w_{3\times3\times3}(\mathbf{C}_l). \tag{4.4}$$

4.2.3 Implementation details

The *U-VixLSTM* was trained and validated using the *Synapse*, *ISIC* and *ACDC* datasets described in Sec. 1.6.1. The preprocessing steps for the *Synapse* and *ACDC* datasets are the same as discussed in Sec. 2.3.2. The dermascopic images and their corresponding masks in the *ISIC* datasets were normalized to the pixel values $\in [0, 1]$. The training data was augmented by rotation and random cropping transformation. The hybrid Dice and Categorical Cross Entropy loss [eqn. (3.11)] was used to train the model.

4.2.4 Results and discussion

Table 4.1 presents an ablation study on *Synapse* data, demonstrating the impact of different numbers (L) of ViL blocks and a varying number of convolution layers along the feature extraction path. The average scores of DSC [eq. (1.21)], $HD95$ [eq. (1.22)], and IoU [eq. (1.20)] are reported for the different organs. The ablation study reflects the trade-off between the complexity of the model and its performance. The best DSC and IoU values were observed for six ViL blocks, and are marked in bold. Performance degraded when the number of

TABLE 4.1: Comparison of different variants of *U-VixLSTM* with increasing number of ViL blocks ($\times L$) and convolution layers, on *Synapse* data.

Ablations	**Model Variants**	**mDSC**	**mIoU**	**mHD95**
# ViL blocks	x 3	0.8118	0.7156	5.56
	x 6	**0.8318**	**0.7323**	4.80
	x 12	0.8289	0.7286	8.57
	x 18	0.8201	0.7189	**4.34**
	x 24	0.8299	0.7280	19.05
# Convolution Layers	3	0.8143	0.7114	5.48
	4	**0.8318**	**0.7323**	**4.80**
	5	0.8314	0.7315	4.85

TABLE 4.2: Comparison with SOTA on multi-organ segmentation (Synapse) dataset.

Model	**DSC**										**IoU**	**HD95**
	Spleen	**Right kidney**	**Left kidney**	**Gall Bladder**	**Liver**	**Pancreas**	**Stomach**	**Right Adrenal**	**Left Adrenal**	**Mean**	**Mean**	**Mean**
U-Net	0.9112	0.9007	0.9181	0.5645	0.9572	0.6967	0.7800	0.6160	0.5872	0.7702	0.6667	38.83
V-Net	0.8874	0.9251	0.9244	0.5858	0.9487	0.7511	0.7953	0.6004	0.4497	0.7631	0.6537	25.29
U-Net++	0.9118	0.9196	0.8905	0.6921	0.9524	0.7536	0.7869	0.5245	0.1032	0.7261	0.6254	55.27
Attention *U*-Net	0.9109	0.8770	0.8720	0.5835	0.9585	0.5566	0.7846	0.5991	0.5398	0.7424	0.6262	51.61
U-Net + EfficientNet-b0	0.8541	0.8919	0.8804	0.6370	0.9077	0.5018	0.7097	0.6048	0.5195	0.7230	0.6056	67.47
U-Net + EfficientNet-b1	0.7414	0.8059	0.8256	0.5500	0.9019	0.5416	0.6160	0.5250	0.5035	0.6679	0.5539	75.85
U-Net 3+	0.8736	0.9229	0.8584	0.6659	0.9383	0.6823	0.7259	0.5657	0.4764	0.7455	0.6254	54.43
Swin UNETR	0.9482	0.9300	0.9245	0.7617	0.9622	0.8046	0.8059	**0.6890**	0.6159	0.8269	0.7261	13.99
UNETR	0.8951	0.9055	0.8950	0.5274	0.9461	0.6668	0.7847	0.5268	0.4978	0.7384	0.6284	26.94
TransUNet	0.8520	0.8828	0.7850	0.7608	0.5870	0.7218	**0.8773**	0.5876	0.4672	0.7246	0.6679	36.81
TransAttUNet	0.9045	0.8761	0.9160	0.4958	0.9408	0.5997	0.6682	0.6559	0.5948	0.7391	0.6580	15.82
Swin-UMamba	0.8778	0.9037	0.9112	0.4206	0.9161	0.6238	0.6527	0.5890	0.5617	0.7174	0.6334	20.10
UNETR++	0.8061	0.8050	0.8201	0.5208	0.8868	0.5249	0.5576	0.5890	0.4577	0.6631	0.5340	8.61
DS-TransUNet	0.4965	0.6113	0.5252	0.5397	0.6790	0.5295	0.5857	0.5745	0.5266	0.5631	0.4644	14.79
U-VixLSTM	**0.9500**	**0.9371**	**0.9366**	**0.8104**	**0.9635**	**0.7878**	0.8304	0.6709	**0.6458**	**0.8318**	**0.7286**	**4.80**

ViL blocks was increased beyond six, indicating overfitting. When the number of parameters increased due to the addition of ViL blocks, the model appeared to memorize the training data instead of learning generalizable features. The six-block configuration appeared to be sufficient on this dataset to capture the global context without the risk of overfitting. Similarly, increasing the number of convolution layers from three to four showed a significant gain in metric scores. This indicates the importance of hierarchical depth in extracting rich representations. However, the performance saturated with a further increase in the number of layers, indicating minimal benefits with an increase in the computational load. The two-part ablation study empirically justifies the design choice of *U-VixLSTM* to hierarchically extract robust spatial features, followed by modeling of global context. This approach efficiently mitigates the risk of overfitting while ensuring computational feasibility.

The performance of our *U-VixLSTM* was next compared with that of the state-of-the-art (SOTA) algorithms in the literature, in terms of metrics DSC, IoU, $HD95$ of eqns. (1.20)-(1.22), on *Synapse*, *ISIC* and *ACDC* datasets. Table 4.2 presents a comprehensive display of the performance on different organs (in *Synapse* data) in the context of DSC. The average results for $HD95$ and IoU are provided for the nine abdominal organs, with the best scores marked in bold. The *U-VixLSTM* was found to outperform other SOTA, with respect to the mean DSC, IoU and $HD95$, with scores of 83.18%, 72.86%, and 4.8,

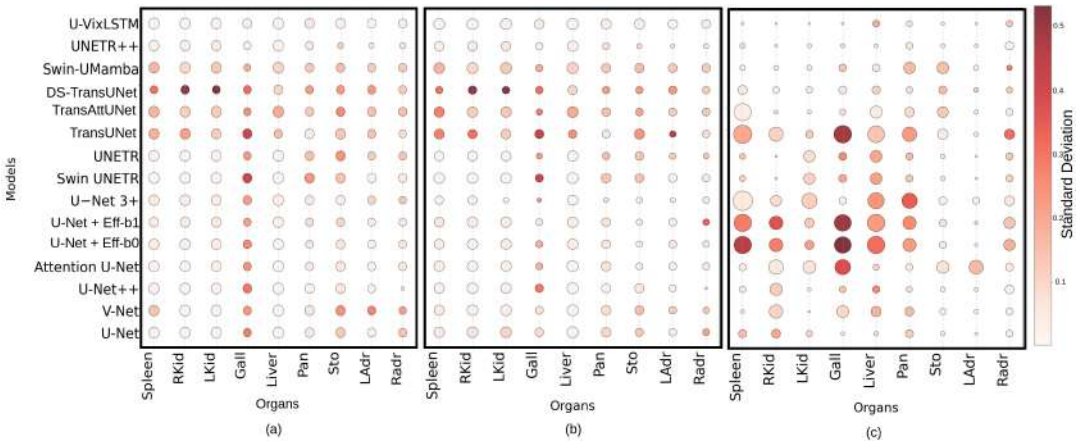

FIGURE 4.2: Dot plot of (a) *DSC*, (b) *IoU*, and (c) *HD*95 metrics for evaluating the performance of *U-VixLSTM* against other baselines on *Synapse* dataset.

respectively. Our model demonstrated superior performance in generating the highest *DSC* values for the segmentation of larger organs (such as the spleen, liver) and smaller organs (such as the kidneys, pancreas, gall bladder and left adrenal gland). Performance remained stable, despite the reduction in organ size, as indicated by the consistently high *DSC* scores observed in smaller and larger organs. This illustrates the robustness of our model in utilizing acquired knowledge on anatomical structures with varying shapes and sizes. It is found to precisely identify and define the target areas possessing unique structures.

Fig. 4.2 visualizes performance consistency across different metrics for *U-VixLSTM* and baselines concerning different abdominal organs. The visualization represents the metric values (denoted by the dot radius) and consistency (represented by the color intensity of the dots). The *U-VixLSTM* exhibits consistent superiority across various organs, as evidenced by the larger radii and uniformly pale color intensities of the dots in Fig. 4.2(a) and (b). This signifies higher accuracy with low standard deviation. Therefore, the high accuracy of the proposed model is both consistent and reliable. Conversely, Swin-UMamba, DS-TransUNet, TransAttUNet and TransUNet, having darker and smaller dots, signifies high variability and lesser accuracy than *U-VixLSTM*. Fig. 4.2(c) represents the *HD*95 metric values across all the organs for different models. A smaller dot radius means higher accuracy in delineating the boundaries for this metric. *U-VixLSTM* consistently demonstrates smallest dot radii with pale color intensities for nearly all the organs. This indicates the superior boundary delineating capabilities compared to baselines, like *U*-Net 3+, *U*-Net with EfficientNet backbone, Attention *U*-Net and TransUNet, having significantly larger and darker dots. While TransUNet attains the highest mean *DSC* value for the stomach, it exhibits a considerably greater standard deviation and higher *HD*95 values in comparison to *U-VixLSTM*. This indicates that while TransUNet effectively segments the majority of the stomach region, it encounters difficulties in accurately delineating organ boundary with consistent performance.

Fig. 4.3 shows the sample segmentation output of *U-VixLSTM* and other architectures, such as Swin UNETR, UNETR, *V*-Net, and TransAttUNet, on

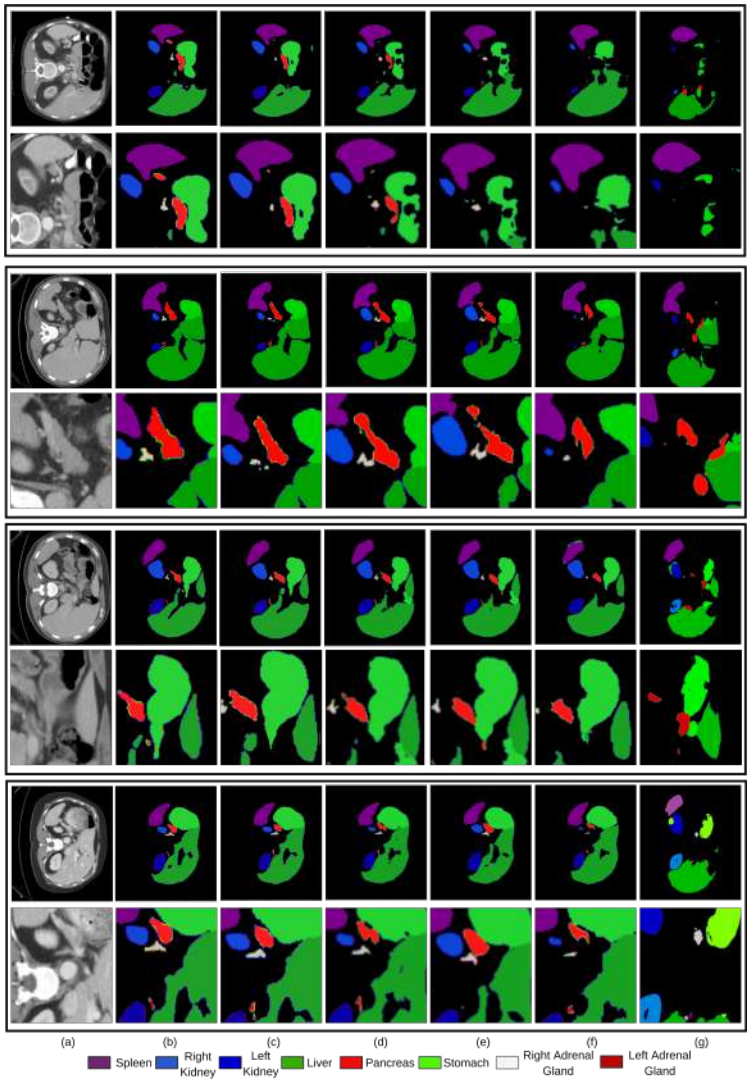

FIGURE 4.3: Comparative performance of *U-VixLSTM* with other baseline architectures, on the *Synapse* dataset, through sample segmentation maps. The first row in each block represents a sample CT slice. The second row provides zoomed-in boxes for a magnified view of specific regions. (a) Input CT image, (b) corresponding ground truth, with the respective output from (c) *U-VixLSTM*, (d) Swin UNETR, (e) UNETR, (f) *V*-Net, and (g) TransAttUNet.

the *Synapse* dataset. It is evident that the *U-VixLSTM* model demonstrates a significantly higher level of similarity to the ground truth, compared to the maps produced by these other baseline models. The results of the Swin UNETR and UNETR models show FP regions corresponding to the stomach, in the second row of Figs. 4.3(d)-(e). The results of V-Net demonstrate a lack of precision in identifying the anatomical regions associated with the adrenal gland, kidneys, pancreas and stomach. The segmentation maps obtained from TransAttUNet exhibit their limited capacity to learn complex representations corresponding to multiple organs, with varying shapes and sizes. The hybrid architecture of *U-VixLSTM* shows effectiveness in segmenting abdominal organs with diverse

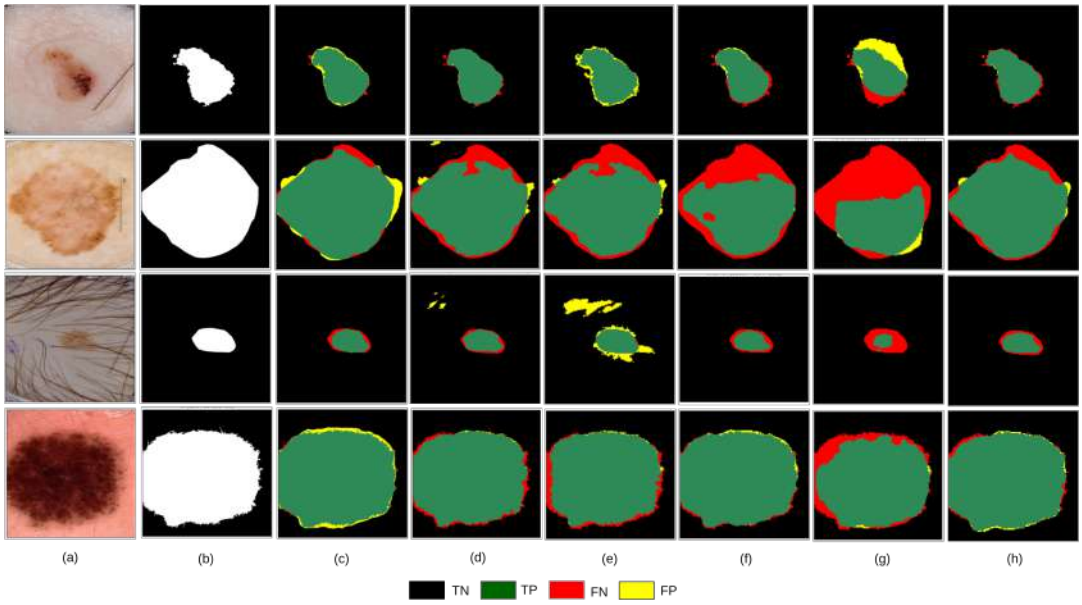

FIGURE 4.4: Comparative performance of *U-VixLSTM* and other baseline architectures, on the *ISIC* dataset, through sample segmentation maps. (a) Input dermoscopic image, (b) corresponding ground truth, with the respective output from (c) *U-VixLSTM*, (d) Swin UNETR, (e) UNETR, (f) *U*-Net 3+, (g) TransAttUNet, and (h) DS-TransUNet.

TABLE 4.3: Comparison with state-of-the-art models on the *ISIC* dataset, with best results marked in **bold**.

Model	DSC	IOU	HD95
$U-$Net	0.8065	0.7197	74.72
Attention $U-$Net	0.7798	0.6845	87.13
$U-$Net + EfficientNet-b0	0.8222	0.7315	12.32
$U-$Net + EfficientNet-b1	0.8120	0.7176	13.44
$U-$Net ++	0.8074	0.7057	13.37
LB-UNet	0.8092	0.7164	57.05
$U-$Net 3+	0.7732	0.6772	14.77
Swin UNETR	0.8187	0.7289	61.06
TransUNet	0.7012	0.6086	90.32
UNETR	0.7601	0.6495	19.04
TransAttUNet	0.7580	0.6610	82.22
DS-TransUNet	0.8212	0.7304	12.50
Swin-UMamba	0.8237	0.7330	56.41
Swin-UMamba†	0.8264	0.7334	55.05
U-VixLSTM	**0.8500**	**0.7611**	**11.31**

shapes and sizes. The CNN layers effectively capture intricate details in high-resolution maps, which is essential for accurately delineating small-sized organs. The ViL block strongly models the global context, making it well-suited for the segmentation of larger organs.

Fig. 4.4 represents the sample segmentation output, as generated by *U-VixLSTM* and other baseline models on the *ISIC* dataset. The prediction made by our *U-VixLSTM* exhibits higher accuracy and precision in relation to the ground truth masks, compared to the maps generated by the other baseline

TABLE 4.4: Comparison with state-of-the-art models on the *ACDC* dataset, with best results marked in **bold**.

Model	DSC				mIOU	mHD95
	LVentricle	RVentricle	Myocardium	Mean		
U-Net	0.7869	0.7956	0.8910	0.8245	0.7130	7.29
V-Net	0.7273	0.6969	0.8108	0.7450	0.6145	6.34
U-Net++	0.6155	0.6828	0.8302	0.7095	0.5693	54.35
Attention *U*-Net	0.7087	0.7637	0.8689	0.7804	0.6568	8.32
U-Net + EfficientNet-b0 backbone	0.6236	0.5314	0.7069	0.6206	0.5219	104.65
U-Net + EfficientNet-b1 backbone	0.6543	0.5217	0.7299	0.6353	0.5026	98.27
U-Net 3+	0.6903	0.6412	0.7355	0.6890	0.5463	149.64
Swin UNETR	0.8059	0.7741	0.8838	0.8213	0.7076	6.02
UNETR	0.7052	0.6845	0.8316	0.7404	0.6049	9.84
TransUNet	0.6398	0.7692	0.8714	0.7601	0.6581	6.93
TransAttUNet	0.6338	0.7420	0.8819	0.7526	0.6432	95.35
Swin-UMamba	0.6640	0.7907	0.8699	0.7749	0.6720	5.45
DS-TransUNet	0.6396	0.7688	0.8654	0.7579	0.6540	5.32
UNETR++	0.7026	0.6977	0.8264	0.7422	0.6121	6.59
U-VixLSTM	**0.8680**	**0.8345**	**0.9104**	**0.8710**	**0.7770**	**5.07**

models. The sample outputs from the other SOTA contain mostly the undersegmented or oversegmented regions. For example, along rows 2 and 3, the baseline models exhibit a significant amount of FN and FP regions; our *U-VixLSTM*, on the other hand, has a visually higher proportion of TP regions and fewer FP and FN pixels. This indicates that these other models had difficulty accurately identifying the boundaries of the target region(s). The quantitative results, presented in Table 4.3, demonstrate the highest values of DSC and IoU along with the lowest $HD95$ metric score for our *U-VixLSTM*. This corroborates the qualitative observation of the enhanced segmentation accuracy of our model.

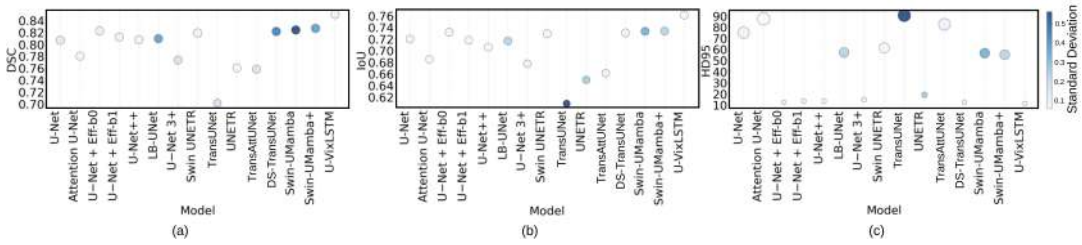

FIGURE 4.5: Dot plot of (a) DSC, (b) IoU, and (c) $HD95$ metrics for evaluating the performance of *U-VixLSTM* against other baselines on *ISIC* data.

Fig. 4.5 provides a detailed analysis of *U-VixLSTM* against the baselines across all the metrics. The pale blue hue of the dot corresponding to *U-VixLSTM* across all the subplots signifies a minimal standard deviation. This demonstrates the superior performance of the model across the diverse dermoscopic images. Precisely delineating the boundaries of the affected region is a significant challenge in skin lesion analysis. Fig. 4.5(c) demonstrates the efficacy of *U-VixLSTM* in addressing this challenge, as indicated by the smallest dot representing the minimum $HD95$ value of 11.31. However, Transformer or Mamba-based models exhibit a larger radius (suboptimal boundary delineation), with darker blue hues signifying discrepancies in boundary prediction.

Table 4.4 presents a comprehensive analysis of performance in different organs in the dataset *ACDC* concerning the *DSC* value. The average results for HD95 and IoU are tabulated for various cardiac organs. *U-VixLSTM* has outperformed baselines, attaining a *DSC* value of 86.8%, 83.45% and 91.04% for

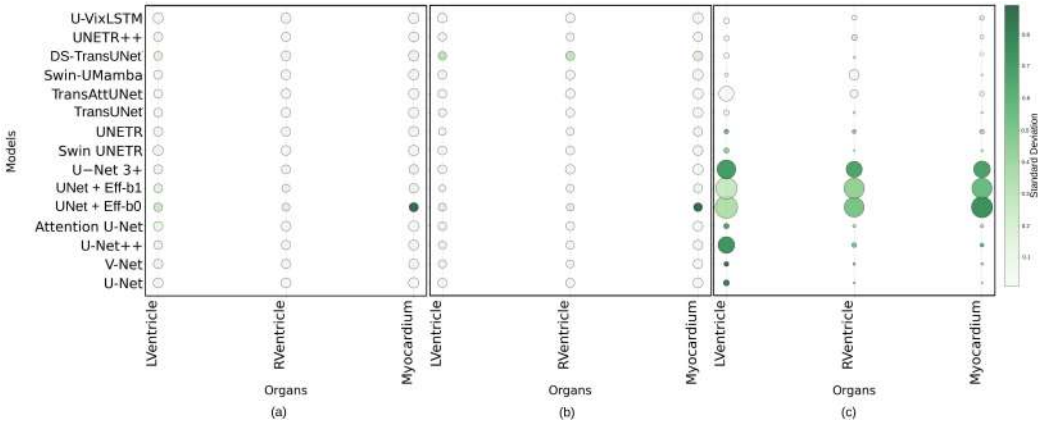

FIGURE 4.6: Dot plot of (a) *DSC*, (b) *IoU*, and (c) *HD*95 metrics for evaluating the performance of *U-VixLSTM* against other baselines on *ACDC* data.

the three distinct cardiac organs, respectively. The mean IoU is 77.7%, exceeding the best baseline by approximately 6%. The dotplot in Fig. 4.6 provides a detailed analysis of the quantitative results. *U-VixLSTM* consistently displayed dots with larger radii and pale color intensities across all the cardiac components in Fig. 4.6(a) and (b). This illustrates balanced and high accuracy across the multiple interconnected cardiac structures. Other approaches, such as DS-TransUNet, Swin U-Mamba, TransUNet and Swin UNETR demonstrated comparable or lower *HD*95 value, as evidenced from Fig. 4.6(c). However, their performance in terms of *DSC* and *IoU* was either significantly lower, as shown in Table 4.4 and Fig. 4.6(a) and (b) or had a higher standard deviation in *HD*95 [Fig. 4.6(c)] compared to *U-VixLSTM*. This suggests *U-VixLSTM* achieves improved and consistent performance in both overlap-based and boundary-based metrics. The ViL block at the bottleneck of *U-VixLSTM* efficiently captured the global view of the cardiac anatomy. This intermediate output guided the feature reconstruction path to generate accurate output regarding boundary delineation and region overlap.

Fig. 4.7 shows sample segmentation maps generated by *U-VixLSTM* along with various baseline methods. The visual representation illustrates the higher-quality outputs generated by our methods relative to established baselines. Instances of mis-segmentations and false positive pixels are evident in the sample maps from Swin UNETR, UNETR, and *V*-Net, as illustrated in Fig. 4.7(d)-(g).

Fig. 4.8 illustrates sample feature maps obtained from the convolution and ViL modules within the feature extraction pathway. The maps visualize the representations acquired by the two modules of our network in relation to the target structures, as specified in the ground truth. The shades of red represent the regions of strongest activation where the model allocates the majority of its attention; whereas, the shades of dark blue signify the lowest activation levels. This visualization helps us observe the areas in the input image that are the most important to the model to define the target structure.

The enhanced performance of *U-VixLSTM* in various modalities is due to its effective capture of local and contextual information. The CNN blocks initially

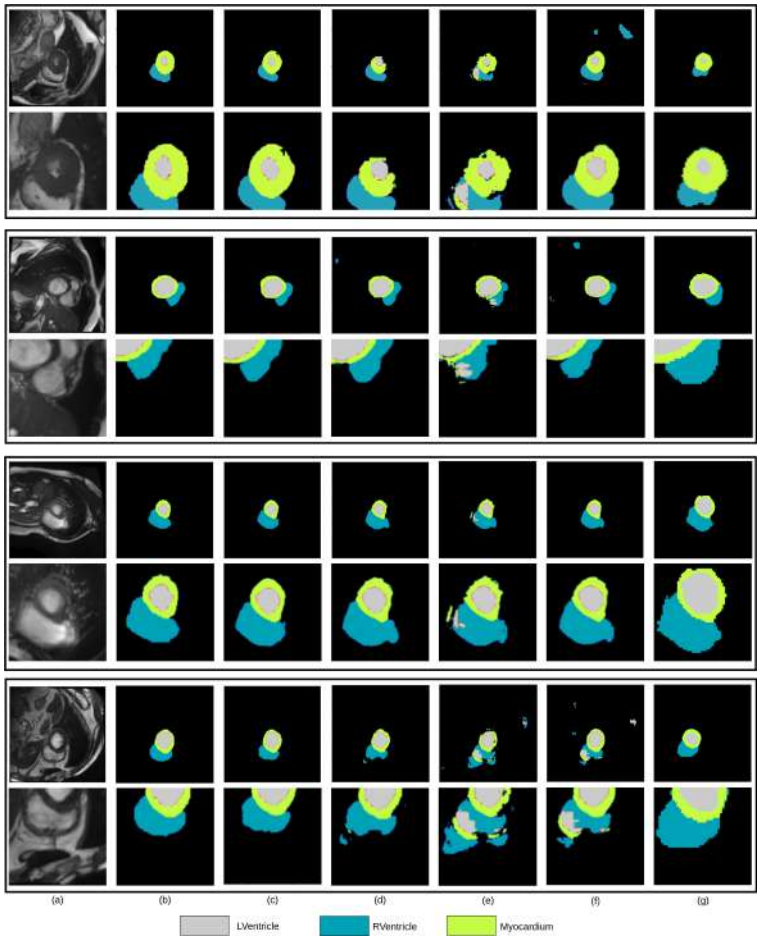

FIGURE 4.7: Comparative performance of *U-VixLSTM* and other baseline architectures, on the *ACDC* dataset, through sample segmentation maps. The first row in each block represents a sample CT slice. The second row in each block provides zoomed-in boxes to provide a magnified view of specific regions. The (a) input CT image, (b) corresponding ground truth, with the respective output from (c) *U-VixLSTM*, (d) Swin UNETR, (e) UNETR, (f) *V*-Net, and (g) TransAttUNet

hierarchically extract fine-grained details, such as edges and local patterns. The ViL blocks in the bottleneck connect distant sections of intermediate feature maps, effectively capturing the general relationships and dependencies between various parts of target structures that have different shapes and sizes. This effectively constructs global contextual information. Incorporating skip connections that merge feature maps from the extraction path with the corresponding levels of the reconstruction path facilitates the localization of anatomical structures. Consequently, ViL enhances the feature extraction process by acquiring superior feature representations while maintaining lower computational costs relative to baseline methods. The gating mechanism of ViL selectively updates the memory matrix to store sharp transitions corresponding to the organ boundaries. This explains the superiority of *U-VixLSTM* in efficiently modelling the boundaries of diverse anatomical structures in comparison to Transformer-based

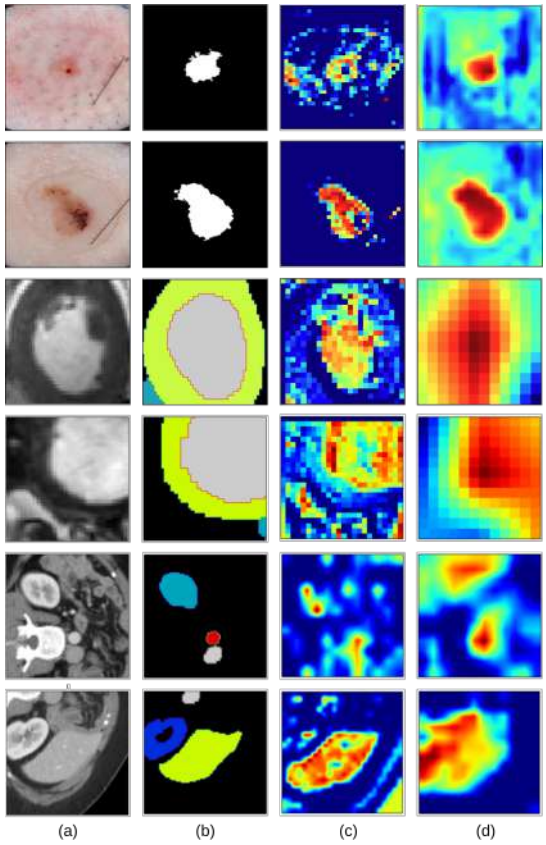

FIGURE 4.8: Visualization of (a) input image, (b) ground truth, and feature maps from (c) CNN and (d) ViL blocks of the feature extraction path.

models. Transformer computes the contextual embedding of each patch using a weighted average of all the image patches. Although being able to effectively capture the global context, the averaging step smoothens the fine-grained details necessary for capturing the boundaries.

Fig. 4.9 provides a graphical analysis of the parameter count, Tera FLoating point OPerations (TFLOPs), and model size on disk, of our *U-VixLSTM* compared to that of other baseline architectures under consideration. The analysis shows that *U-VixLSTM* has the lowest number of parameters and TFLOPs (floating point operations per second) compared to the other SOTA. This reiterates the claim about the superior computational efficiency of our model. Hence, it demonstrates potential for deployment in resource-constrained environments.

4.3 Rotational U-Vision-xLSTM

The *U-VixLSTM* illustrated the efficacy of ViL as a powerful backbone for the medical image segmentation task. Although ViL successfully captured the global context, a key challenge still existed; especially, for volumetric inputs. The anatomical structures of interest span across different slices, owing to which it is difficult to comprehend the overall complex structure. This requires a

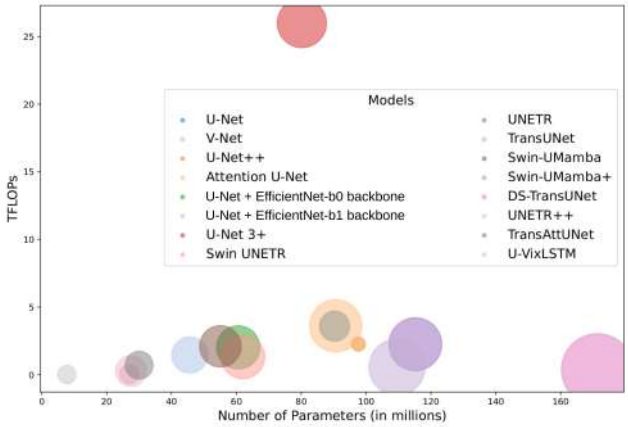

FIGURE 4.9: Comparison with SOTA with respect to number of parameters (in millions), TFLOPs, and model size on disk (in MB). Bubble size is indicative of model size.

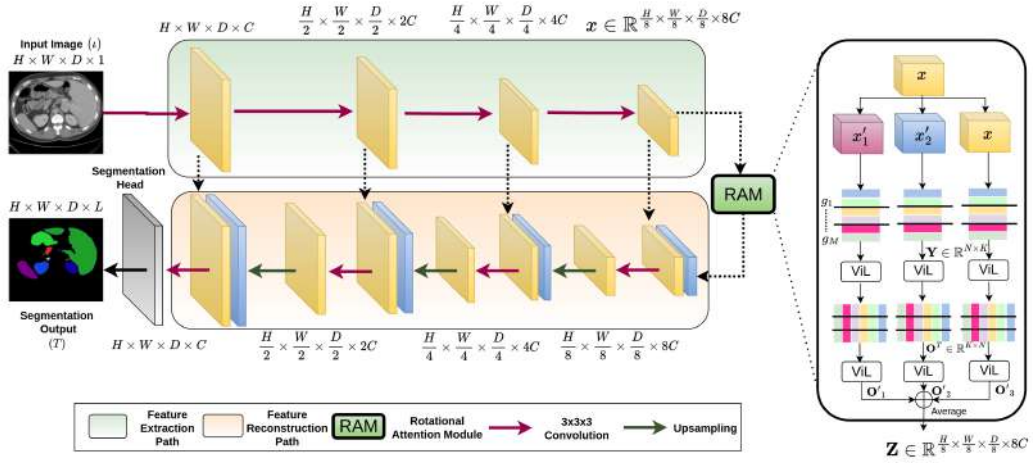

FIGURE 4.10: Architectural framework of *Rot-UViL*.

comprehensive contextual understanding to capture the relationships between different axes of the entire data volume. To address this, an advanced version of *U-VixLSTM* called *Rotational U-VixLSTM (Rot-UViL)* was developed to model crucial cross-dimensional dependencies. *Rotational Attention Module (RAM)* is the core module of this model. It processes feature map volumes from multiple perspectives to build a rich and robust representation of target regions.

The organizational framework of *Rot-UViL* is illustrated in Fig. 4.10. The three main components of *Rot-UViL* include the feature extraction path, the Rotational Attention Module *(RAM)*, and the feature reconstruction path. As in *U-VixLSTM*, the input image I is hierarchically processed by the feature extraction path consisting of a series of convolution layers to produce the intermediate output $x \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times \frac{D}{8} \times 8C}$. The bottlenck consists of the *RAM* block, which captures the cross-dimensional dependencies within x. Following this lies the feature reconstruction path, which hierarchically upsamples the *RAM*

output to generate the final segmentation output. As in *U-VixLSTM*, skip-connections combine the feature maps from the feature extraction path with the corresponding stages in the feature reconstruction path. This ensures that fine-grained details, captured in early stages, are propagated to deeper levels for precise segmentation.

4.3.1 Rotational Attention Module (RAM)

Here x is routed to three different pathways for parallel processing. One path captures spatial attention within x, while the other two simultaneously compute the interactions between C and the H or W axes. This helps capture the cross-dimensional dependencies between spatial dimensions H, W and C. The first path models the correlation between the C and H axes. Keeping the D dimension fixed, x is rotated 90° anticlockwise along the H axis to obtain $x_1' \in \mathbb{R}^{\frac{H}{8}\times 8C\times\frac{D}{8}\times\frac{W}{8}}$. Similarly, the second path computes the relationship between the C and W axes by rotating x 90° anticlockwise along the W axis. The resultant tensor x_2' has dimensions $\mathbb{R}^{8C\times\frac{W}{8}\times\frac{D}{8}\times\frac{H}{8}}$. Here, x is processed without any rotation along the last path to model global spatial dependencies throughout the input.

Lemma 3 (Norm preservation property of *RAM* module)**.** *The rotation operation on $x \in \mathbb{R}^{C\times H\times W\times D}$ in RAM module to produce new volume x' is an isometry which preserves the Frobenius norm $||x||_F = ||x'||_F$.*

Proof. Frobenius norm of x is given by

$$||x||_F = \sqrt{\sum_{i=1}^{C}\sum_{j=1}^{H}\sum_{k=1}^{W}\sum_{l=1}^{D}|x_{i,j,k,l}|^2}. \tag{4.5}$$

Frobenius norm of x_1' (by rotating x along H axis) is given by

$$||x_1'||_F = \sqrt{\sum_{j=1}^{H}\sum_{i=1}^{C}\sum_{l=1}^{D}\sum_{k=1}^{W}|x_{j,i,l,k}|^2}. \tag{4.6}$$

Since rotation of x along H axis is a permutation of the indices

$$\implies \{|x_{i,j,k,l}|^2\} = \{|x_{j,i,l,k}|^2\}. \tag{4.7}$$

$\therefore$ Using eqns. (5.6)-(5.8) we have

$$\begin{aligned} ||x||_F &= \sqrt{\sum_{j=1}^{H}\sum_{i=1}^{C}\sum_{l=1}^{D}\sum_{k=1}^{W}|x_{j,i,l,k}|^2} \\ &= ||x_1'||_F. \end{aligned} \tag{4.8}$$

Similarly, $||x||_F = ||x_2'||_F$. $\therefore$ By transitivity we have $||x||_F = ||x_1'||_F = ||x_2'||_F$. □

The above property indicates that the activation is not affected by the rotation operation, ensuring numerical stability in *RAM* during training. Applying rotations to x, while keeping D fixed, helps the network to proficiently acquire channel-height and channel-width correlation within each 2D slice of volumetric data. This complements the global 3D analysis in the third pathway; thereby, allowing the model to capture both local in-plane features and global cross-plane dependencies for superior representation learning.

The tensors along each path are divided into non-overlapping patches of dimension $p \times p \times p$. The patches are then flattened to 1D vectors and projected onto a $K-$dimensional embedding space, with learnable positional embeddings, to yield $Y \in \mathbb{R}^{N \times K}$. Here, N denotes the total number of patches, each row in Y represents a patch, and each column indicates a feature channel.

Y is divided along the rows into M groups. Each group is represented as $[g_i]_{i=1}^{M} \in \frac{N}{M} \times K$, as illustrated in Fig. 4.10. The ViL blocks model the interactions among the different patches in each group, with $O_i \in \mathbb{R}^{\frac{N}{M} \times K}$ being the output obtained from ViL. This is expressed as

$$O_i = \text{ViL}(g_i). \tag{4.9}$$

This approach helps to learn local patterns and finer details within nearby patches. The output of the concurrent processing of each group, by ViL, is aggregated to generate $O \in \mathbb{R}^{N \times K}$. Then O is transposed such that the columns in O^T represent the different patches and the rows denote the feature channels. Next, O^T is divided into groups along rows; with each group processed concurrently by the ViL blocks. This helps capture global contextual information across all patches.

The parallel processing of groups reduces computational complexity (due to smaller matrix multiplications), enhances memory efficiency on GPUs, and potentially improves performance through superior cache utilization with increased parallelization. The aggregation of resultant tensors from the three parallel paths can be mathematically expressed as

$$Z = \frac{1}{3}(O'_1 + O'_2 + O'_3). \tag{4.10}$$

Here, O'_i represents the output along the i th path and $Z \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times \frac{D}{8} \times 8C}$ is the final output of the *RAM* module.

4.3.2 Implementation details

The *Rot-UViL* was trained and validated using the *Synapse* and *AdrenalSeg* datasets of Sec. 1.6.1. The preprocessing steps for the *Synapse* and *AdrenalSeg* datasets are the same as discussed in Sec. 2.3.2. The hybrid Dice and Categorical Cross Entropy loss [of eqn. (3.11)] were used to train the model.

TABLE 4.5: Ablation study for different group sizes with respect to *DSC* and GPU memory usage.

Group Size	DSC	GPU Memory Usage (in MB)
1	0.7713	6.81
2	0.857	0.57
4	0.836	0.29
8	0.8372	0.15
16	0.7877	0.09

TABLE 4.6: Comparison of *Rot-UViL* with baseline models on *Synapse* and *AdrenalSeg*.

Model	Synapse												AdrenalSeg		
	DSC										Mean IoU	Mean HD95	DSC	IoU	HD95
	Spleen	Right kidney	Left kidney	Gall Bladder	Liver	Pancreas	Stomach	Right Adrenal	Left Adrenal	Mean					
U-Net	0.91	0.90	0.92	0.56	0.96	0.70	0.78	0.61	0.59	0.77	0.67	38.83	0.68	0.71	60.22
V-Net	0.89	0.92	0.92	0.59	0.95	0.75	0.80	0.60	0.45	0.76	0.65	25.29	0.68	0.69	56.49
U-Net++	0.92	0.91	0.89	0.69	0.95	0.75	0.79	0.52	0.10	0.73	0.62	55.27	0.63	0.50	72.59
Attention *U*-Net	0.91	0.88	0.88	0.58	0.96	0.55	0.79	0.60	0.54	0.74	0.63	51.61	0.70	0.57	81.82
U−Net with EfficientNet-b0	0.85	0.90	0.88	0.64	0.90	0.50	0.71	0.60	0.52	0.72	0.60	67.47	0.65	0.52	56.38
U−Net with EfficientNet-b1	0.74	0.80	0.83	0.55	0.90	0.54	0.61	0.52	0.50	0.67	0.55	75.85	0.61	0.52	95.61
U−Net 3+	0.87	0.92	0.86	0.67	0.94	0.69	0.73	0.57	0.48	0.74	0.62	54.43	0.69	0.56	48.01
Swin UNETR	0.95	0.93	0.92	**0.76**	0.96	0.80	0.80	0.69	0.62	0.83	0.73	13.99	0.71	0.58	51.19
UNETR	0.89	0.90	0.89	0.53	0.95	0.67	0.79	0.53	0.50	0.74	0.63	26.94	0.60	0.48	100.95
TransUNet	0.85	0.89	0.79	0.76	0.59	0.72	0.88	0.59	0.47	0.72	0.67	36.81	0.73	0.60	34.79
TransAttUNet	0.90	0.87	0.91	0.49	0.94	0.59	0.66	0.65	0.59	0.73	0.65	15.81	0.67	0.56	108.00
DS-TransUNet	0.49	0.61	0.52	0.53	0.67	0.52	0.58	0.57	0.53	0.56	0.57	27.58	0.56	0.45	**19.70**
Swin UMamba	0.87	0.90	0.91	0.42	0.91	0.62	0.66	0.59	0.56	0.72	0.63	20.1	0.64	0.55	71.13
Rot-UViL	**0.96**	**0.95**	**0.94**	0.71	**0.97**	**0.83**	**0.84**	**0.70**	**0.70**	**0.85**	**0.75**	**12.75**	**0.86**	**0.77**	30.08

4.3.3 Results and discussion

Table 4.5 depicts the effect of group size on *DSC* and GPU memory utilization. Grouping substantially improved performance, thus demonstrating its effectiveness in capturing relevant patterns. A significant increase of $\sim 8\%$ in the *DSC* was observed when the group size was increased to two from one (no grouping). This improvement in segmentation accuracy was also accompanied by a drastic reduction in the GPU memory usage by more than ten times (from 6.81 to 0.57 MB). This approach is memory efficient, with a drastic decrease in GPU memory consumption occurring with increasing group size. Thus, it becomes advantageous in resource-constrained environments.

However, there was a trade-off between the expanded group size and the *DSC* score value. Although GPU memory consumption reduced by a substantial amount with larger group sizes (reaching a minimum of 0.09 MB with a group size of 16), segmentation accuracy decreased significantly. This is because the number of groups is inversely proportional to the length of token sequence processed by the ViL block. The ViL processes a longer token sequence with smaller number of groups, which efficiently captures meaningful long-range dependencies. In contrast, a larger number of groups result in shorter token sequences, which inhibits the model from capturing these global relationships, leading to a drop in *DSC* despite significant gains in terms of GPU memory efficiency. Therefore, a group size of two was chosen for our experiments.

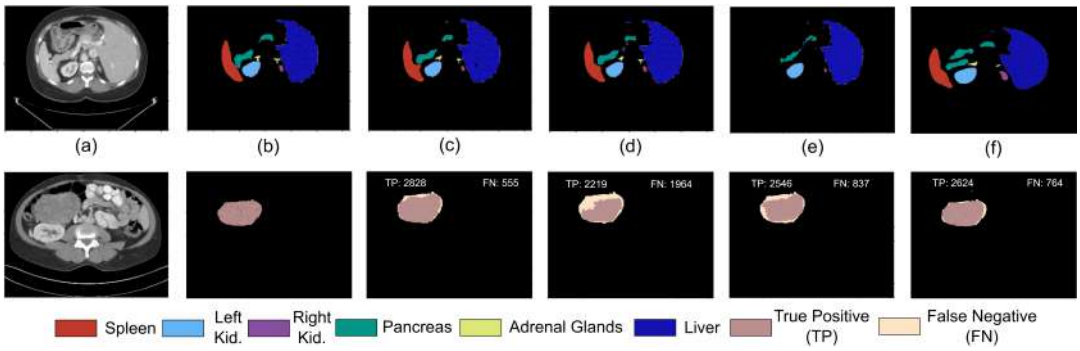

FIGURE 4.11: Sample segmentation maps on *Synapse* (first row) and *AdrenalSeg* dataset (second row): (a) Input CT slice, (b) Ground Truth, with output from (c) *Rot-UViL*, (d) Swin UNETR, (e) UNETR, and (f) TransAttUNet.

Table 4.6 presents a detailed analysis of the quantitative performance of *Rot-UViL*, compared to several popular baselines, on the *Synapse* and *AdrenalSeg* datasets. *Rot-UViL* outperformed the baselines by achieving the highest mean scores across all the metrics on the *Synapse* dataset. This is a clear improvement compared to the second-best model Swin UNETR. The *Rot-UViL* efficiently segmented the smaller and complex organs, *viz.* the pancreas (DSC of 0.83) and the adrenal glands (DSC of 0.70). Its superior performance in delineating smaller (challenging) organs is due to the novel *RAM* module, which efficiently captures the fine-grained cross-dimensional dependencies for a comprehensive representation. However, certain baselines exhibited superior performance in specific organs, *viz.* TransUNet and Swin UNETR achieved the highest DSC values for the stomach and gall bladder, respectively. Nevertheless, the consistent performance of *Rot-UViL*, especially in smaller challenging organs, shows the versatility of the developed model.

Fig. 4.11 presents a qualitative analysis of the segmentation results on the *Synapse* dataset. The visual results complement the findings of the quantitative analysis, establishing the superiority of *Rot-UViL* compared to other baselines. The segmentation output of *Rot-UViL* [Fig. 4.11(c)] closely aligns with the ground truth [Fig. 4.11(b)], where it efficiently delineates the complex boundaries of both large and small organs. In contrast, baselines such as UNETR do not appropriately identify the spleen, pancreas, and adrenal glands. The TransAttUNet incorrectly segments the pancreas. Although Swin UNETR successfully delineates most organs, it does not produce precise boundaries; as evidenced by the over-segmentation of the pancreas in Fig. 4.11(d).

The developed model also exhibited superior performance with respect to the values of DSC and IoU, on the *AdrenalSeg* dataset, as observed from Table 4.6. In fact, *Rot-UViL* outperformed the best baseline by $\sim 13\%$ in the DSC value. However, DS-TransUNet produced the best performance for the $HD95$ metric. This indicates that while *Rot-UViL* excels in identifying the overall tumor region, it exhibits reduced precision while delineating the finer boundaries of the tumor. However, it exhibited comparatively better boundary segmentation performance compared to other hybrid or purely CNN-based approaches.

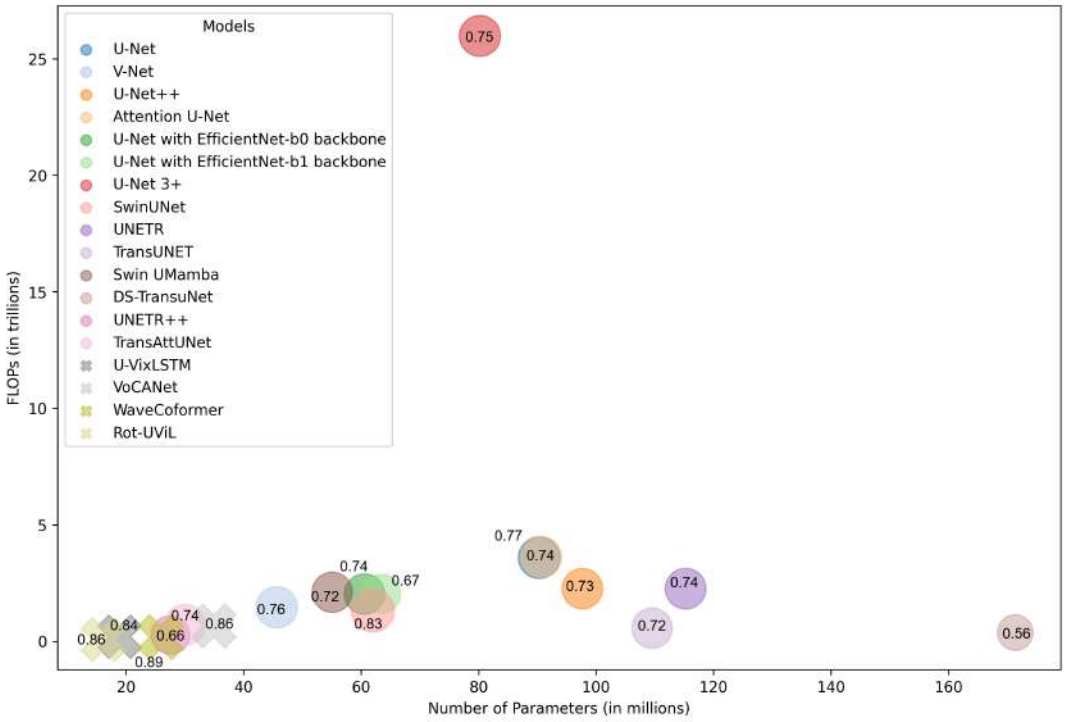

FIGURE 4.12: Trade-off analysis of the developed architectures versus state-of-the-art baselines. The x-axis represents the number of parameters in millions, the y-axis denotes the computational computing overhead (TFLOPs), and the bubble diameter corresponds to the segmentation performance (Mean DSC) for the *Synapse* dataset.

The significant gains in correctly identifying the overall tumor region represent its overall superiority and robustness. The second row in Fig. 4.11 visually confirms the quantitative superiority of *Rot-UViL*. The model achieved the highest TP and lowest FN pixel count among the baselines. In contrast, the compared models exhibited significant under-segmentation, with Swin UNETR failing to identify a major portion of the tumor with an FN count of 1964. Similarly, UNETR and TransAttUNet missed significant boundary regions of the tumour.

The consistent superior performance of *Rot-UViL* on the diverse set of multi-organ and tumor segmentation tasks can be attributed to its architectural design. The feature-extraction path captures the low-level features and finer textural details using CNNs, which helps in the precise delineation of complex boundaries. The novel RAM module is the pivotal component of the architecture with its ability to model cross-dimensional dependencies within the CNN output. Such comprehensive, multi-faceted understanding allows *Rot-UViL* to learn an advanced contextual representation of the target, leading to high segmentation accuracy across both smaller organs and larger irregular tumors.

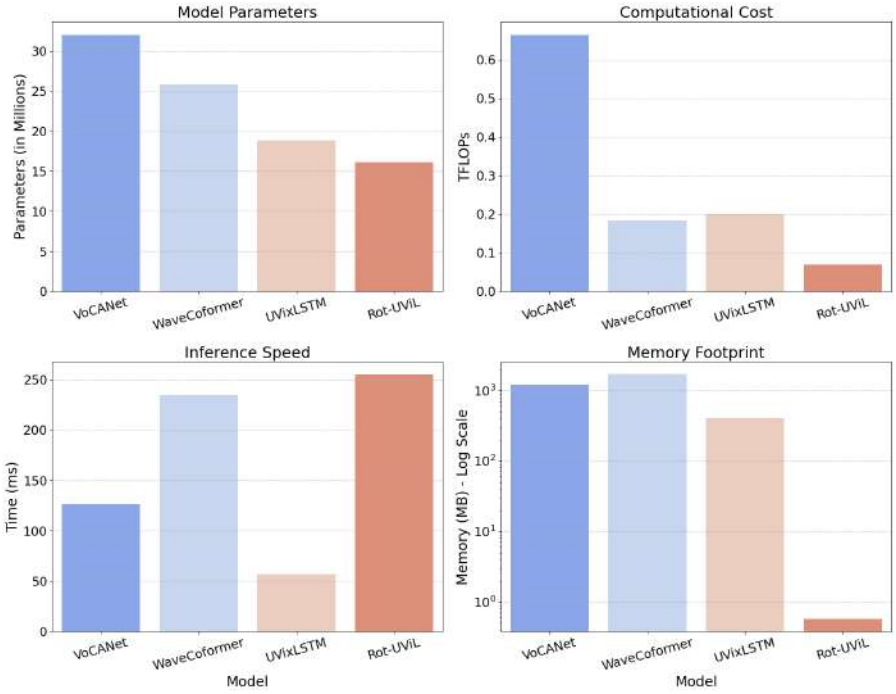

FIGURE 4.13: Comparative analysis of computational and memory costs of the developed models.

4.4 Comparison between Models Developed

A detailed comparative analysis among the four models developed so far, presented in Figs. 4.12 and 4.13, demonstrate a trade-off between segmentation accuracy and computational efficiency. Traditional baselines exhibit a compromise between the segmentation performance and computational efficiency, as seen in Fig. 4.12. Standard convolutional networks (e.g., *U*-Net 3+) have massive computational overhead (up to 25.98 TFLOPs) for moderate accuracy, while heavy transformer-based models (e.g., DS-TransUNet and TransUNet) consume extensive memory (exceeding 100M-170M parameters). In contrast, the novel models developed throughout the thesis (marked by x in Fig. 4.12) shift the performance-efficiency balance towards the bottom-left quadrant. The *Rot-UViL* exhibits the highest average segmentation accuracy in different abdominal organs in the *Synapse* dataset. The highest performance, combined with the lowest parameter count and computational cost (as evident from Fig. 4.13), makes it well-suited for memory-constrained environments, despite its slower inference speed. The *WaveCoformer* is a strong competitor with the second-best mean *DSC*. However, its high memory demand (Fig. 4.13) makes it suitable for offline medical image analysis, where computational resources are abundant. Additionally, modeling of frequency information is crucial for high accuracy while capturing subtle details. On the other hand, *U-VixLSTM* has the fastest inference speed, as seen in Fig. 4.13. This makes it well-suited for applications requiring real-time feedback, where both speed and precision are important. In spite of a promising segmentation performance by the fully

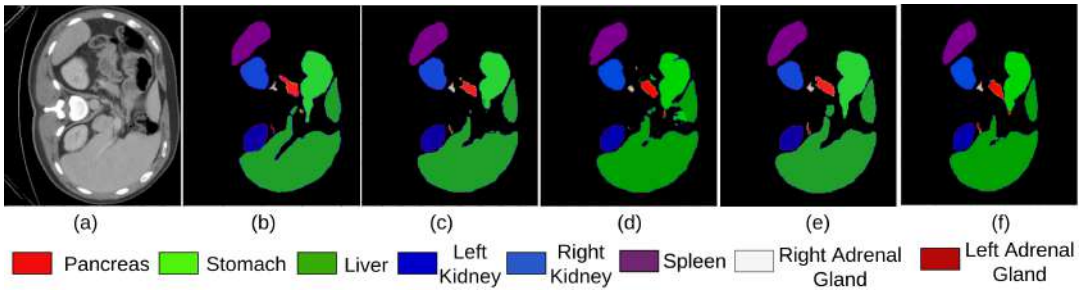

FIGURE 4.14: Comparative analysis of qualitative results of the developed models on *Synapse*, with (a) Input CT slice, (b) Ground truth, output from (c) *WaveCoformer*, (d) *VoCANet* (e) *U-VixLSTM* and (f) *Rot-UViL*.

convolutional *VoCANet*, it was surpassed by the hybrid architectures in terms of resource efficiency. The qualitative results, illustrated in Fig. 4.14, visually validate the findings of the quantitative analysis.

TABLE 4.7: Global statistical ranking of the evaluated architectures based on the Friedman test. A lower mean rank indicates consistently higher segmentation performance globally.

Overall Rank	Architecture	Mean Statistical Rank
1	**Rot-UViL**	**1.67**
2	**WaveCoformer**	**2.56**
3	***U*-VixLSTM**	**3.44**
4	Swin UNETR	3.56
5	**VoCANet**	**5.83**
6	*U*-Net	8.33
7	*V*-Net	8.89
8	*U*-Net++	9.56
9	Attention *U*-Net	10.11
10	TransAttUNet	10.72
11	TransUNet	11.22
12	UNETR	11.22
13	*U*-Net 3+	11.44
14	Swin UMamba	12.11
15	*U*-Net with EfficientNet-b0	12.22
16	*U*-Net with EfficientNet-b1	14.89
17	DS-TransUNet	15.22

Non-parametric statistical analysis was conducted to validate the comparative performance of the developed models. The Friedman test was done to evaluate the global statistical significance among the developed models and baselines. The models were ranked based on their segmentation performance across different anatomical structures in the *Synapse* dataset. As shown in Table 4.7, the developed models have the highest rank with *Rot-UViL* securing the best global mean rank. Additionally, a pairwise Wilcoxon test was employed to have a head-to-head comparison between the different models and baselines, as seen in Table 4.8. A p-value of < 0.05 mathematically confirms that the segmentation performance of the developed network is significantly superior to the compared baseline.

TABLE 4.8: Pairwise Wilcoxon signed-rank test p-values comparing the proposed architectures against baseline models.

Baselines	p-**value**			
	vs. VoCANet	**vs. U-VixLSTM**	**vs. WaveCoformer**	**vs. Rot-UViL**
U-Net	< 0.05	< 0.05	< 0.05	< 0.05
V-Net	< 0.05	< 0.05	< 0.05	< 0.05
U-Net++	< 0.05	< 0.05	< 0.05	< 0.05
Attention U-Net	< 0.05	< 0.05	< 0.05	< 0.05
U-Net with EfficientNet-b0	< 0.05	< 0.05	< 0.05	< 0.05
U-Net with EfficientNet-b1	< 0.05	< 0.05	< 0.05	< 0.05
U-Net 3+	< 0.05	< 0.05	< 0.05	< 0.05
UNETR	< 0.05	< 0.05	< 0.05	< 0.05
TransAttUNet	< 0.05	< 0.05	< 0.05	< 0.05
TransUNet	< 0.05	< 0.05	< 0.05	< 0.05
Swin UMamba	< 0.05	< 0.05	< 0.05	< 0.05
DS-TransUNet	< 0.05	< 0.05	< 0.05	< 0.05
Swin UNETR	< 0.05	< 0.05	< 0.05	< 0.05

4.4.1 Formal computational complexity analysis

This section presents the formal computational complexity analysis of the architectures developed in this thesis. This analysis mathematically quantifies the computational overhead associated with the key mechanism of each model.

Computational complexity of *VoCANet*:

The encoder-decoder backbone of *VoCANet* is made up of depthwise separable convolution kernels whose compute time can be expressed as $\mathcal{O}(NC+NC^2)$. Here, $N = H \times W \times D$ and C represents the number of feature channels. The channel attention within the global context attention module comprises dilated convolution kernels (dil), global average pooling (GAP) and MLP to generate weights for channel-wise recalibration. The complexity of the channel attention can be expressed as

$$Cost_{cha} = Cost_{dil}+Cost_{GAP}+Cost_{MLP} = \mathcal{O}(NC^2)+\mathcal{O}(NC)+\mathcal{O}(C^2) \approx \mathcal{O}(NC^2).$$

The complexity of the spatial attention path, made up of $3\times3\times3$ and pointwise convolution kernels along with trilinear upsampling is

$$Cost_{spa} = \mathcal{O}(NC^2) + \mathcal{O}(NC) + \mathcal{O}(N) \approx \mathcal{O}(NC^2).$$

$\therefore$ Total cost of *VoCANet* becomes $\mathcal{O}(NC + NC^2) + \mathcal{O}(NC^2) + \mathcal{O}(NC^2) \approx \mathcal{O}(NC^2)$.

Computational complexity of *WaveCoformer*:

Each stage of *WaveCoformer* processes the input tensor volume in parallel across two paths – one dedicated to wavelet domain analysis and the other to spatial domain processing. The *SpectraConv* module translates the input tensor to the wavelet domain using DWT and processes the output sub-bands using dilated and depthwise separable convolution, as described in Sec. 3.2.1. The

computational complexity of *SpectraConv* is

$$Cost_{spectraconv} = \mathcal{O}(C.N) + \mathcal{O}(C^2.\frac{N}{64}).$$

Parallel to *SpectraConv*, the *Dual Attention* module first identifies relevant channels within the input tensor, followed by modeling global context using windowed self-attention mechanism. The detailed steps are described in Sec. 3.2.1. The computational complexity of this path is defined as

$$Cost_{dual} = Cost_{cha} + Cost_{wsa} = \mathcal{O}(NC) + \mathcal{O}(\frac{C^2}{r}) + \mathcal{O}(NC^2).$$

Next, the *CCA* module, comprising 3D convolution kernels, combines the output of these parallel branches to generate robust feature representations. The compute time can be expressed as $\mathcal{O}(NC^2)$. $\therefore$ the total computational cost of *WaveCoformer* can be mathematically represented as

$$Total = Cost_{spectraconv} + Cost_{dual} + Cost_{cca} \approx \mathcal{O}(NC^2).$$

Computational complexity of *U-VixLSTM* **and** *Rot-UViL*:

U-VixLSTM is primarily driven by Vision-xLSTM blocks, which use the mLSTM mechanism. Unlike self-attention in ViTs, which computes a quadratic attention matrix, mLSTM sequentially processes the tokens. Let T be the total number of embedded patches, defined as $T = \frac{H \times W \times D}{P^3}$. Here, H, W, D, P represent the height, width, depth, and patch dimension. Let d be the embedding dimension of each patch. At every time step t, the mLSTM computes Q, K, and V projections along with updating the covariance matrix C_t. These operations scale quadratically with the embedding dimension, bounded by $\mathcal{O}(d^2)$. Since the model sequentially processes T patches, the total computational complexity of the ViL block can be expressed as $\mathcal{O}(T.d^2)$.

The rotational attention module in Rot-UViL divides the patch matrix of dimension $T \times d$ into M groups. Therefore, the total computational complexity becomes $M.\mathcal{O}(\frac{T}{M}d^2) = \mathcal{O}(Td^2)$.

4.5 Conclusion

This chapter introduced Vision-xLSTM-based architectures, *viz. U-VixLSTM* and *Rot-UViL*, to address the huge computing requirements of the hybrid CC-Transformer based segmentation scenarios. The *U-VixLSTM* integrated Vision-xLSTM (ViL) into the $U-$Net, demonstrating the efficacy of the linear computational complexity of ViL over the quadratic computational complexity of Vision Transformers. The model outperformed state-of-the-art approaches in precise boundary delineation, along with the fastest inference speed. Subsequently, the *U-VixLSTM* was extended to the *Rot-UViL* to accurately delineate

complex anatomical structures that span over multiple slices of an input image volume. The Rotational Attention Module within the *Rot-UViL* modeled cross-dimensional dependencies by processing volumetric tensors from multiple perspectives. The *Rot-UViL* outperformed baselines with respect to segmentation accuracy and computational efficiency. Finally, a detailed evaluation of the different models presented in this thesis helped validate the trade-off in performance-efficiency among these architectures.

The findings state the nature of the application of the developed models in real-world clinical settings. The impressive performance of *Rot-UViL*, especially its low memory footprint, makes it ideal for deployment in resource-constrained platforms, such as edge devices or smartphone applications – with limited access to powerful GPUs. Although the inference speed is slow, it serves as a scalable AI solution. The *U-VixLSTM*, with its low latency, is suited for time sensitive applications like rapid patient screening. The ViL-based models demonstrate that increasing segmentation accuracy does not directly imply higher computational burden. They offer powerful solutions in a wide range of clinical settings.

While developing efficient deep architectures is a significant step towards resource-efficient image analysis, another promising direction lies in focusing computational load on the relevant regions of interest. The processing of redundant information leads to a waste of computational cycles. The subsequent chapter explores a dynamic adaptation of computations, in the relevant regions of interest, guided by some auxiliary information such as spatial prompts during inference. This approach reduces the computational burden of ViT-based models during inference; thereby, achieving runtime efficiency along with dynamic input adaptability.

Chapter 5

Spatially-Aware Token Processing in Efficient Vision Transformers

"Perfection is achieved, not when there is nothing more to add, but when there is nothing left to take away."

— Antoine de Saint-Exupéry, Airman's Odyssey

5.1 Introduction

The development of computationally efficient models follows two broad strategies. The first is the design of inherently computationally efficient models, as discussed in the preceding chapters. The alternative path to achieve computational efficiency is through algorithmic optimization. This approach optimizes complex models by dynamically modifying their existing operation.

Model compression techniques, such as pruning, are an effective way of algorithmic optimization. Pruning reduces the computational cost of a model by removing less significant elements from its architecture. Static pruning removes the fixed components in the model architecture, before inference; thereby, irreversibly eliminating the components considered less significant. In contrast, dynamic pruning adaptively adjusts the computation inference by selectively deactivating model components based on input. Dynamic pruning is a preferred choice, over static pruning, due to its ability to adapt to the unique features of each input image. This is a significant advantage over the highly variable nature of medical image data.

This chapter explores this new direction by introducing Prompt-driven Adaptive Token (*PrATo*) pruning [36] to optimize ViT-based state-of-the-art segmentation models. The novel framework targets the processing of redundant information, which leads to the wastage of computational cycles. Many tokens processed by ViTs are semantically redundant. Token pruning increases efficiency by reducing the number of processed tokens. *PrATo* incorporates structured spatial priors to direct token retention. It facilitates input-specific adaptation of the network computation. This guarantees the preservation of tokens essential for depicting the structure of the target object, while redundant or less informative tokens get removed. The proposed method improves the precision of segmentation through a direct incorporation of the spatial constraint, while optimizing computational efficiency. The research contribution is summarized below.

- Design of a novel prompt-driven dynamic token pruning framework to increase the efficiency of ViTs through an integration of auxiliary spatial prompts. This helps to localize relevant regions related to the target structure(s).
- Development of an efficient entropy-based scoring mechanism to quantify the relevance of different image tokens. The parameter-free approach prevents additional computational overhead during training and inference.
- Establishment of generalizability of the strategy over different state-of-the-art ViT-based medical image segmentation models, on different publicly available datasets.

The rest of the chapter is organized as follows. Section 5.2 provides a comprehensive overview of the steps involved in the *PrATo* token pruning framework. The implementation details of the framework, along with qualitative and quantitative results, are presented in Section 5.3. The application of our framework, as embedded in state-of-the-art ViT-based segmentation models, is also described. Finally, Section 5.4 concludes the chapter.

5.2 Prompt-driven Adaptive Token Pruning

This section describes the new *PrATo* framework for integration into ViT blocks of the prevalent ViT-based medical image segmentation models. Recent high-performance medical image segmentation models employ an *U*-shaped encoder-decoder architecture [106]. *PrATo* can be seamlessly integrated at different stages of existing segmentation models, to adaptively remove irrelevant tokens from being processed in subsequent steps of the network.

5.2.1 Token generation in ViT

ViTs transform an input image (or feature map volume) $I \in \mathbb{R}^{C\times H\times W}$ into a sequence of tokens [33]. Here, H, W and C represent the height, width, and channel dimensions of I. The input is partitioned into a set of non-overlapping patches $Z; |Z| = \frac{HW}{P^2}$, with each patch having dimension P^2. They are flattened and linearly projected in C'-dimensional embedding space, resulting in the sequence $P' \in \mathbb{R}^{Z\times C'}$. These embedded vectors are called tokens, with Z denoting the total number of tokens generated. Positional embedding is added to the tokens to retain their spatial locations. The $P_0 \in \mathbb{R}^{Z\times C'}$ is the sequence of position-sensitive tokens that serve as input to a transformer encoder.

The encoder block consists of alternating layers of multihead self-attention units (MSA) and feed-forward network (MLP) units. The MSA captures global contextual relationships between different tokens, performing parallel self-attention across multiple heads, to operate in various representational subspaces. The MLP is independently applied to the output of each head, to enhance their representational capacity by including an additional nonlinear transformation. The operations within a transformer block are expressed as

$$Y_{out} = MLP[LN\{MSA(LN(P_0))\}] + P_0, \tag{5.1}$$

where LN [71] is Layer Normalization and $Y_{out} \in \mathbb{R}^{Z\times C'}$ is the final output of the transformer block.

5.2.2 Prompt-driven adaptive pruning

Prompts serve as a prior by offering high-level, task-specific information [63]. This directs the model to focus on segmenting specific structures within the input image, resulting in effective segmentation. Consequently, it allows token selection based on contextual information. The framework of Fig. 5.1 is referred

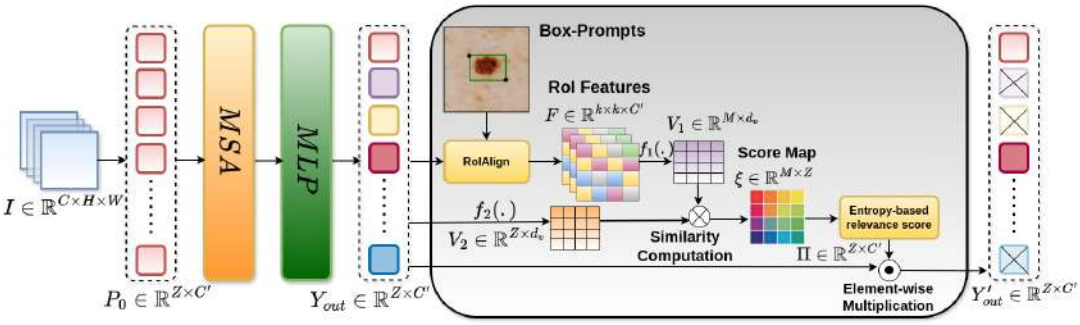

FIGURE 5.1: The *PrATo* framework for pruning ViT tokens.

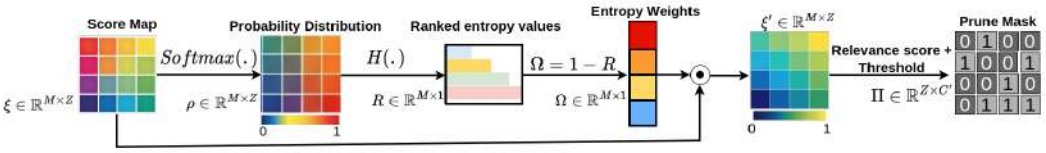

FIGURE 5.2: Computation of pruning mask Π from similarity score map ξ.

to as *prompt-driven adaptive pruning*, where essential tokens are preserved while down-weighting the irrelevant ones to minimize their processing.

The Y_{out} is spatially rearranged to 2D maps and fed to RoIAlign [54] along with box-prompts, to extract region-specific features $F \in \mathbb{R}^{k\times k\times C'}$. The box prompts serve as spatial priors, highlighting the most informative regions within the input. The RoIAlign performs a soft-crop to extract the region-specific feature embeddings corresponding to the RoI defined in the prompt from Y_{out}. Each token in Y_{out} is a contextual embedding of an image patch. Consequently, the ViT token grid may misalign with the box-prompt coordinates making naive selection of tokens by hard cropping infeasible. The RoIAlign ensures a spatially coherent feature representation, unlike naive selection of tokens, to preserve spatial consistency.

Here F is a summarization of the prompted region generated by RoIAlign. By itself it is unable to identify globally relevant ViT tokens for segmenting the target structures. Therefore, a similarity score map $\xi \in \mathbb{R}^{M\times Z}$ is calculated between the region-specific features, encoded by F and Y_{out}, to measure the relevance of each token. The F and Y_{out} are first transformed into lower-dimensional embeddings $V_1 \in \mathbb{R}^{M\times d_V}$ and $V_2 \in \mathbb{R}^{Z\times d_V}$, respectively, with projection matrices $f_1 \in \mathbb{R}^{C'\times d_V}$ and $f_2 \in \mathbb{R}^{C'\times d_V}$. This reduces computational complexity while improving feature discrimination. Here M denotes the number of feature vectors in V_1. The similarity score is computed as

$$\xi(V_1, V_2) = V_1 V_2^T / \sqrt{d_V}. \tag{5.2}$$

The next step involves evaluating the prune mask $\Pi \in \mathbb{R}^{Z\times C'}$ from the similarity score mask ξ to filter out irrelevant tokens, as shown in Fig. 5.2. Here, ξ is normalized along the token dimension to generate a probability distribution $\rho \in \mathbb{R}^{M\times Z}$ by softmax operation. The probability that ρ_{ij} of the ith feature vector V_{1i} of V_1 is relevant to the jth feature vector of V_{2j} of V_2 is given

as $\rho_{ij} = \frac{e^{(\xi_{ij})}}{\sum_{k=1}^{Z} e^{\xi_{ik}}}$, with ξ_{ij} denoting the similarity score between V_{1i} and V_{2j}. The softmax amplifies the higher similarities, while suppressing lower ones, to distinctly identify the confident matches. Shannon entropy $H(.)$ [118] is next calculated for the probability distribution of each V_{1i}, over all ViT tokens, as

$$H(V_{1i}) = -\sum_{j=1}^{Z} \rho_{ij} \log(\rho_{ij}), \tag{5.3}$$

with $H(V_{1i})$ quantifying the uncertainty of association between V_{1i} and the set of all ViT tokens. A lower value of $H(V_{1i})$ implies a strong association with a subset of tokens. On the other hand, a higher value indicates that V_{1i} is uncertain about its association with the tokens. In lieu of adopting thresholding to filter tokens with low similarity, entropy assesses token relevance in a statistically significant manner. This facilitates a confidence-aware selection process, with the framework being interpretable and applicable across various images. Tokens exhibiting high confidence (low entropy) are preserved, while those that are ambiguous (high entropy) are down-weighted.

An inverse entropy weighting scheme is applied to prioritize tokens having lower uncertainty values. The entropy values are first ranked in ascending order. The ranks $R_i | i \in [1, M]$ are normalized to ensure their uniform spacing between 0 and 1. Inverse entropy weights $\Omega_i = 1 - R_i$, with $\Omega = \{\Omega_i | i \in [1, M]\}$, ensure that higher entropy values (more uncertainty) are assigned to lower weights, and vice versa.

Lemma 4 (Uniformity of Weights)**.** *If the normalized ranks R_i are uniformly distributed on $[0, 1]$, then the transformed weights Ω are also uniformly distributed over $[0, 1]$. In other words, applying the transformation $\Omega_i = 1 - R_i$ preserves the uniformity of the distribution, merely reversing the order of values.*

Proof. Since $R_i \sim \text{Uniform}(0, 1)$, its cumulative distribution function (CDF) is

$$F_R(x) = P(R_i \leq x) = x, \quad x \in [0, 1].$$

For the transformed weights $\Omega_i = 1 - R_i$, the CDF becomes

$$F_\Omega(x) = P(\Omega_i \leq x) = P(1 - R_i \leq x) = P(R_i \geq 1 - x).$$

Since $R_i \sim \text{Uniform}(0, 1)$, we substitute its CDF as

$$P(R_i \geq 1 - x) = 1 - P(R_i \leq 1 - x) = 1 - (1 - x) = x.$$

Thus,

$$F_\Omega(x) = x, \quad x \in [0, 1].$$

Since this matches the CDF of a uniform distribution over $[0, 1]$, we conclude that $\Omega_i \sim \text{Uniform}(0, 1)$. □

The weighted similarity score map becomes $\tilde{\xi} = \xi.\Omega$, with the overall relevance r_i of the ith token being the mean of the weighted similarity scores

$$r_i = \frac{1}{C'} \sum_{j=1}^{C'} \tilde{\xi}_{ij}. \tag{5.4}$$

The final step modifies the original tokens, based on their relevance scores. A binary mask $\Pi \in \{0,1\}^Z$ is generated by applying a threshold T on the relevance scores as

$$\Pi_i = \begin{cases} 1, & \text{if } \sigma(r_i) > T, \\ 0, & \text{otherwise.} \end{cases} \tag{5.5}$$

Here σ is the sigmoid function to squash the relevance scores to range $[0,1]$, and Π is applied to ViT tokens Y_{out} to effectively remove those that are deemed irrelevant based on our prompt-guided weighting and thresholding mechanism. The masked token Y'_{out} is generated by element-wise multiplication ($\odot$) of Π with Y_{out}, such that $Y'_{out} = \beta \odot Y_{out}$.

Lemma 5 (Retaining high-relevance token set)**.** *If $H(V_{1i})$ is low, there exists a typical set of high-relevance tokens S_T which concentrates the effective information capacity. For the threshold value T, we have $S_T \subseteq S_P$, where S_P is the final set of tokens retained.*

Proof. Since $H(V_{1i})$ is low, it satisfies $H(V_{1i}) \leq \log_2(Z) - \Delta$, for some token reduction factor $\Delta > 0$.
By the fundamental property of typical sets,
$\exists\ S_T \subset Z$ such that $S_T \leq Z.2^{-\Delta}$ and S_T concentrates the effective information capacity as

$$\sum_{k \in S_T} \rho_{ik} \geq 1 - \epsilon; for\ \epsilon > 0.$$

$\therefore$ For any token $k \in S_T$ and $l \notin S_T$, we have

$$\rho_{ik} >> \rho_{il}. \tag{5.6}$$

The weights Ω constitute a monotonically decreasing function of $H(.)$ with a low $H(.)$ corresponding to a high Ω. Additionally, ρ_{ik} is a monotonically increasing function of the similarity score ξ_{ik}.
$\therefore$ From eqn. (5.6) $\implies S_{ik} >> S_{il}$.
For important tokens $i \in S_T$ and unimportant tokens $j \notin S_T$, we get

$$r_i >> r_j. \tag{5.7}$$

Thus, the algorithm assigns higher scores to important tokens and lower scores to the unimportant ones. Since there exists a significant gap between the relevance scores, $\exists\, r_{min} = min_{i \in S_T}\{r_i\}$ and $r_{max} = max_{j \notin S_T}\{r_j\}$.
By eqn. (5.7), we have

$$r_{min} > r_{max}.$$

A token k is retained if $\sigma(r_k) > T$.
$\therefore r_k > \sigma^{-1}(T)$.

TABLE 5.1: Component ablation of *PrATo* on *ACDC* dataset.

Variants	DSC			mIoU	mHD95
	Ventricle		Myocardium		
	Left	Right			
w/o Entropy	0.4923	0.6819	0.7586	0.5393	18.48
Naive Cropping	0.5272	0.7068	0.7739	0.5692	**15.89**
PrATo	**0.6146**	**0.7171**	**0.819**	**0.6129**	19.87

$\implies r_{min} > \sigma^{-1}(T) > r_{max}$.
$\forall i \in S_T$, the relevance score $r_i > \sigma^{-1}(T)$ which ensures that every token in S_T is retained.
Thus, $S_T \subseteq S_P$. □

Corollary 1. *Given the token reduction factor* $\Delta > 0$*, then* $|S_T| < Z$.

Proof. By the fundamental property of typical sets we have

$$|S_T| \leq Z.2^{-\Delta}. \tag{5.8}$$

From the corollary we get

$$\begin{aligned} \Delta > 0 &\implies 2^{\Delta} > 2^{0}, \\ &\implies 2^{\Delta} > 1, \\ &\implies 2^{-\Delta} < 1, \\ &\implies Z.2^{-\Delta} < Z. \end{aligned}$$

∴ From eqn. (5.8) we get

$$|S_T| \leq Z.2^{-\Delta} < Z \implies |S_T| < Z. \tag{5.9}$$

□

5.3 Implementation and Results

The *PrATo* framework was evaluated on state-of-the-art ViT-based segmentation models, *viz.* UNETR, TransUNET, SegFormer and Swin UNETR. The models were trained using the *ISIC* and *ACDC* datasets of Sec. 1.6.1. The preprocessing steps for the *ACDC* dataset were the same as discussed in Sec. 2.3.2. The dermoscopic images were processed as discussed in Sec. 4.2.3. The hybrid Dice and Categorical Cross Entropy loss [eqn. (3.11)] was used to train the models.

5.3.1 Ablation study

The contribution of the core components of *PrATo* is explored in Table 5.1. Two of its variants were compared with our proposed framework. These are (i)

TABLE 5.2: Ablation of *PrATo* on *ACDC*, over threshold T.

Threshold		**GFLOPs**	**mDSC**
Fixed	0.1	15.25	0.6131
	0.2	15.78	0.6425
	0.3	13.32	0.6701
	0.5	9.07	0.6612
Percentile	$25th$	26.78	0.6933
	$50th$	37.31	0.6952
	$75th$	18.66	0.6575

TABLE 5.3: Ablation study for spatial dimension of RoI Align (k) on *ACDC* dataset.

k =	3	5	7	9
mDSC	0.6761	**0.7169**	0.6405	0.6627

without entropy-based weighting, and (ii) replacing RoIAlign with Naive cropping. The second variant directly selected the tokens, with center coordinates located within the box prompt. Our *PrATo* framework attained the highest *DSC* value among all cardiac organs, and exhibited superior average *IoU* with respect to the other variants. This indicates that the integration of RoIAlign for feature extraction and entropy-based token scoring could effectively identify the relevant tokens. The variant lacking the entropy-weighting scheme demonstrated a notable decrease in *DSC* [eqn. (1.21)] and *mIoU* [eqn. (1.20)] values. It exclusively employed the unrefined features of RoIAlign to identify important tokens. Therefore, the entropy-weighting scheme was found to be necessary to filter ambiguous features, leading to precise token selection. Naive cropping indiscriminately selected all tokens within the box, as it lacked the advanced feature extraction capabilities of RoIAlign. This can be evidenced by the significant drop in *DSC* values obtained for the left ventricle and myocardium. Therefore, RoIAlign was the superior feature extraction approach for this task.

However, the average HD95 value ($mHD95$) [eqn. (1.22)] was lower with the naive cropping approach than with the proposed PrATo framework. The fine-grained details, especially around the edges, were smoothed out by directly selecting all the underlying tokens within the box-prompt. The spatially averaged query directed the similarity mechanism to select contiguous groups of tokens; thereby, leading to predictions characterized by simpler boundaries. Despite reduced penalties for smoother boundaries, as indicated by the $HD95$ metric, the overall shape of the target structure was not accurately predicted. This is evident from the lower values of *DSC* and *IoU*.

Table 5.2 quantifies the experimental results at fixed threshold values (0.1, 0.2, 0.3, 0.5) and adaptive percentile-based threshold values ($25th$, $50th$, $75th$) for T. Lower values of T, *viz.* 0.1 and 0.2, exhibited a reduced value of mean *DSC* over the different classes in *ACDC*. This suggests that allowing low-relevance token processing tends to introduce noise, which impacts the final output quality. Although increasing the threshold value to 0.5 led to a significant decrease in GFLOPs, it also resulted in a slight drop in segmentation

accuracy, indicative of over-pruning.

In contrast, percentile-based thresholding had higher segmentation accuracy as evident from higher $mDSC$ values. Lower percentile values ($25th$) and ($50th$) retained comparatively more tokens than the higher percentiles ($75th$), as indicated by the increase in Giga Floating Point Operations (GFLOPs). The T at the $25th$ percentile is found to have an optimal balance between accuracy and computational efficiency. Discarding 75% of total tokens resulted in suboptimal segmentation performance due to information loss. Adaptive percentile-based thresholding provided greater flexibility to the token pruning mechanism as compared to fixed threshold values, and achieved a superior balance between segmentation performance and computational costs. The fixed threshold values remained constant over all inputs, making them inflexible to the distinct score distribution of each input. Percentile-based thresholding always preserved a constant proportion of relevant tokens in relation to a specific input. The risk of discarding essential tokens was minimal, thus maintaining a good segmentation performance.

Table 5.3 presents the results of the ablation study analyzing the impact of different spatial dimensions of the region-specific characteristics derived from RoIAlign. Segmentation performance improved significantly from $k = 3$ to 5. The 5×5 feature map obtained, by setting $k = 5$, resulted in a high-dimensional query that allowed the model to learn discriminative and fine-grained details from the bounded region of the box prompt. However, increasing the value of k might have included specific textures or noise unique to the small target region. Therefore, while finding its association with the rest of the tokens, the model failed to generalize and overfitted to the specific input prompt. Consequently, the segmentation performance dropped for higher values of k. Therefore, $k = 5$ was chosen as the final value of the proposed framework.

5.3.2 Comparative study

Table 5.4 provides the experimental findings to understand the potential of the prompt-guided *PrATo*, as compared to the other state-of-the-art pruning frameworks (described in Sec. 1.4.4). It also includes a crucial baseline, called Random Token Masking, which randomly drops a fixed number of tokens without any guidance from spatial priors like the box-prompt. This highlights the impact of pruning without any auxiliary guidance, like spatial priors from box-prompts. Class-wise DSC, along with the average values IoU and $HD95$, are reported for the $ACDC$ dataset. Here *PrATo* exhibited superior performance in terms of DSC, achieving improvements of 7.28%, 3.61% and 1.05% for the left ventricle, right ventricle and myocardium, respectively, when compared to the next best baseline. The mean IoU exceeded that of the second-best method by nearly 7%. Random Token Masking demonstrated a marginal advantage over *PrATo* in terms of $mHD95$; however, its low DSC and $mIoU$ values suggest under-segmentation of regions that align closely with the true organ regions, as corroborated in the visualization of Fig. 5.3(f).

A qualitative comparison is presented in Fig. 5.3 to visually assess segmentation maps from the different state-of-the-art pruning frameworks on the

TABLE 5.4: Comparative analysis of *PrATo*, with other pruning frameworks, on *ACDC* and *ISIC* datasets.

Framework	*ACDC*					*ISIC*			Inference Time (ms)
	DSC			mIoU	mHD95	DSC	IoU	HD95	
	Ventricle		Myocardium						
	Left	Right							
DynamicViT	0.5115	0.5898	0.7302	0.4879	24.03	0.8493	0.7539	21.59	10.11
EvoViT	0.5418	0.681	0.8085	0.5561	24.5	0.8541	0.7566	**8.30**	59.15
DToP	0.4463	0.5598	0.6845	0.4447	21.7	0.8511	0.7449	35.41	15.23
Random Token Masking	0.5094	0.5927	0.7136	0.4905	**18.85**	0.8402	0.7377	13.83	**2.66**
STP	0.3958	0.4224	0.5677	0.3388	21.72	0.8544	0.7483	10.91	13.21
TRAM	0.4867	0.5448	0.6844	0.4562	20.08	0.6424	0.5514	40.02	3.07
PrATo	**0.6146**	**0.7171**	**0.819**	**0.6129**	19.87	**0.8634**	**0.7678**	17.34	3.55

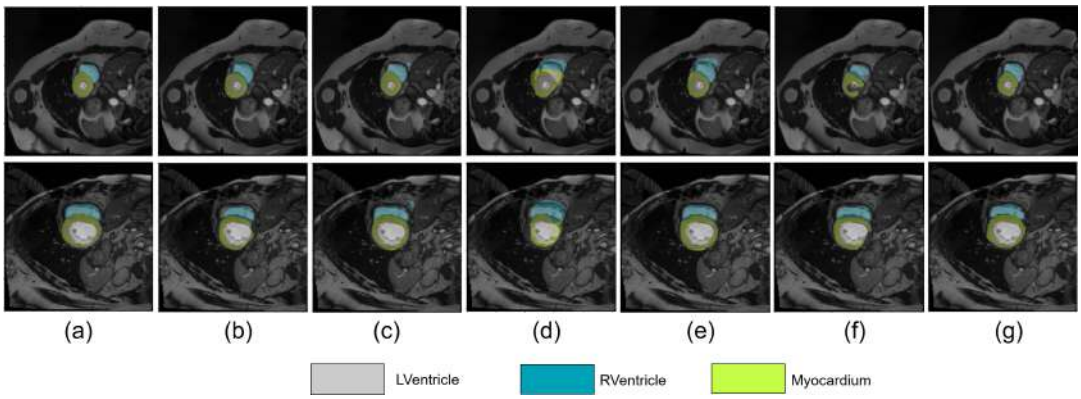

FIGURE 5.3: Sample segmentation maps, comparing *PrATo* with other pruning frameworks, on the *ACDC* dataset. (a) Input MRI image with overlaid ground truth, and sample outputs from (b) Dynamic ViT, (c) EvoViT, (d) STP, (e) DToP, (f) Random Token Masking, and (g) *PrATo* frameworks.

ACDC dataset. The output of *PrATo* was found to be visually closest to the ground truth, as compared to the other approaches. The overall shapes of the target structures were accurately captured, indicating the importance of spatial priors in retaining relevant tokens. The STP produced imprecise output, with inaccurate boundaries and under-segmented regions of the left ventricle and myocardium, respectively [Fig. 5.3(d)]. The Dynamic ViT, EvoViT and DToP over-segmented the left ventricle, as evident from the top row of Fig. 5.3(b), (c), and (e).

Our *PrATo* also achieved the best performance, based on *DSC* and *IoU* scores, on the *ISIC* dataset; as shown in Table 5.4. This indicates that the segmentation output of *PrATO* had the best overlap with the ground truth, accurately delineating the lesion region as compared to the other methods. However, EvoViT attained the best *HD*95 score. This difference in boundary delineation ability, between EvoViT and *PrATo*, originate from their different pruning mechanisms. EvoViT retains irrelevant tokens to maintain the spatial grid structure, thus preserving finer boundary details. Although the box-prompt in *PrATo* served as a robust prior to identifying the overall region, it represents a coarser form of guidance that may cause inaccuracies in boundary delineation.

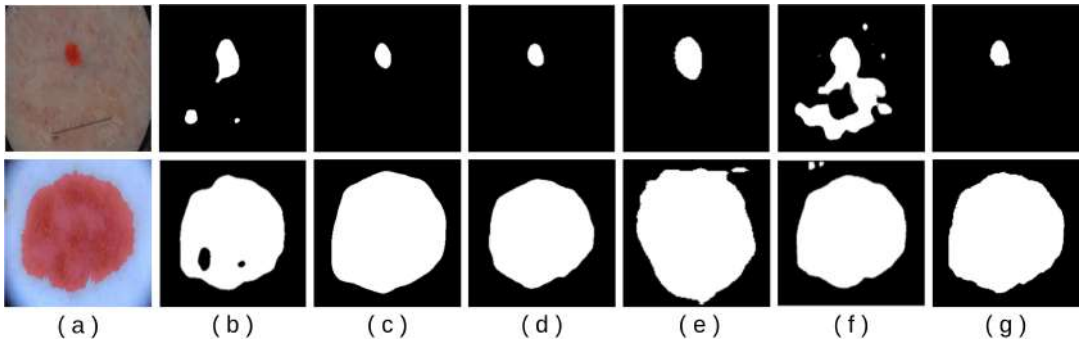

FIGURE 5.4: Sample segmentation maps, comparing *PrATo* with other pruning frameworks on the *ISIC* dataset. (a) Input dermoscopy image with overlaid ground truth, and sample outputs from (b) Random Token Masking, (c) Dynamic ViT, (d) EvoViT, (e) STP, (f) DToP, and (g) *PrATo* frameworks.

Fig. 5.4 illustrates the sample segmentation maps of the *PrATo* and other pruning frameworks in the *ISIC* data set. The proposed framework generated visually precise segmentation output for both small and large lesions. The lesion boundaries were smooth and the overall shape aligned closely with the truth of the ground. Random Token Masking produced several under-segmented regions in the output. STP and DToP display numerous oversegmented areas characterized by irregular boundaries.

The last column of Table 5.4 shows that *PrATo* outperformed several high-accuracy methods, such as EvoViT and Dynamic ViT, in terms of inference speed. It was approximately 3x faster than Dynamic ViT and nearly 15x faster than the high-performing model EvoViT. The ability of *PrATo* to exhibit competitive performance, with reduced inference time, highlights the efficiency of our prompt-guided mechanism. Although Random Token Masking and TRAM were comparatively faster than *PrATo*, their performance degradation makes them unsuitable for clinical use.

5.3.3 Effectiveness analysis

The *PrATo* was next integrated with various ViT-based state-of-the-art segmentation models, to demonstrate its generalizability and model-agnostic nature. Table 5.5 summarizes the values of *DSC* for these segmentation models and their pruned versions using *PrATo*, in the datasets *ACDC* and *ISIC*. The proposed framework was applied to the output of ViT blocks for the non-hierarchical models, such as UNETR and TransUNET, and subsequently to the patch merging stage for hierarchical models like Swin UNETR and SegFormer. This ensured that the architectural operations in the models were not compromised, and the reduced set of tokens was propagated to subsequent stages. UNETR and TransUNET exhibited significant performance gains ($1\% - 7\%$ approx.) in *DSC* values in both datasets. This suggests that the prompt-based spatial prior helps the model focus more on relevant regions. Swin UNETR and SegFormer

TABLE 5.5: Quantitative analysis of the implementation of *PrATo* across various medical image segmentation baseline models on the *ACDC* and *ISIC* datasets. The DSC value is reported for each baseline and their corresponding pruned version.

Models \ Datasets	*ACDC*						*ISIC*	
	Ventricle				Myocardium		Baseline	Pruned
	Left		Right					
	Baseline	Pruned	Baseline	Pruned	Baseline	Pruned		
UNETR	0.5521	0.5943	0.6952	0.7201	0.7773	0.7820	0.8330	0.8754
Swin UNETR	0.6400	0.6621	0.7940	0.7854	0.8462	0.8543	0.8850	0.8832
TransUNET	0.6401	0.7061	0.7693	0.8312	0.8710	0.8821	0.8242	0.8253
SegFormer	0.6841	0.6791	0.8221	0.8163	0.8981	0.9024	0.8121	0.8220

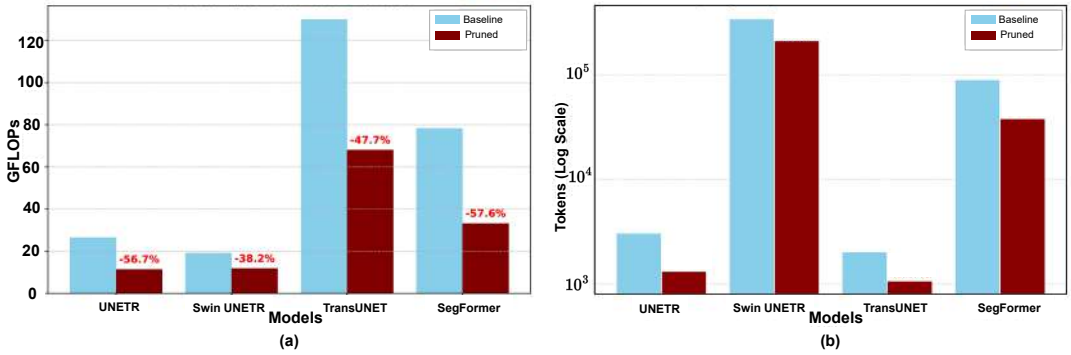

FIGURE 5.5: Graphical plot comparing (a) GFLOPs and (b) Token Sparsity of the segmentation models, with respect to their pruned versions.

demonstrated stable performance, even after substantial token removal, as corroborated from Fig. 5.5. Therefore, our *PrATo* framework improved the overall segmentation accuracy by retaining relevant tokens of target structures.

Models with a full self-attention mechanism, such as UNETR and TransUNET, evaluate the association between all possible pairs of tokens. However, this might include processing irrelevant tokens. *PrATo* functions as a regularizer, enabling the model to focus on relevant tokens corresponding to the target structures. This improves the segmentation performance. Models with local self-attention, like Swin UNETR and SegFormer, are optimized to model the relevant local features; thereby, resulting in efficient segmentation performance. Therefore, the primary benefit of *PrATo*-incorporated variants of such models lies more in their computational savings than in performance improvement.

Fig. 5.5(a) illustrates that our PrATo framework significantly reduces computational costs across all models. The pruned variants of SegFormer and UNETR show the highest reduction in terms of GFLOPs. Fig. 5.5(b) depicts a significant decrease in the token density of the pruned variants, compared to their respective baselines. This elimination of irrelevant tokens leads to computationally efficient segmentation models.

Fig. 5.6 qualitatively analyzes the effectiveness of our framework. The samples highlight cases where the pruned version effectively addresses issues of over- and under-segmentation in the predictions made by the baselines. The

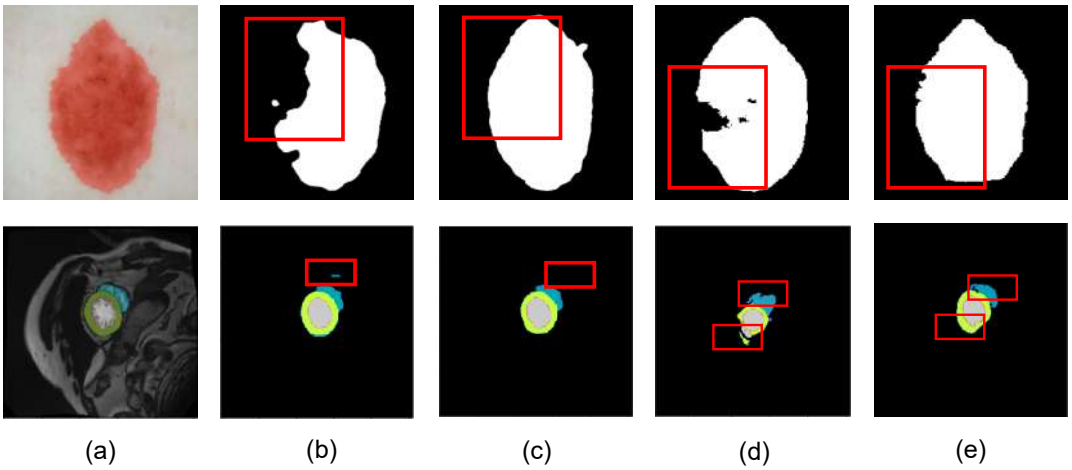

FIGURE 5.6: Qualitative results on sample images, illustrating (a) input image, with overlaid ground truth, and prediction from (b) SegFormer (baseline), (c) SegFormer (pruned), (d) UNETR (baseline), and (e) UNETR (pruned).
Row 1: *ISIC*, Row 2: *ACDC*, sample images.
The red boxes denote the comparison area.

experiments suggest that *PrATo* generates a versatile framework to improve the efficacy of different segmentation models.

Fig. 5.7 visualizes the sample token retention maps from both *ACDC* and *ISIC* datasets, to gain insight into the effectiveness of our approach. The retained tokens (marked by yellow patches) correspond to the relevant spatial regions associated with the target structures. *PrATo* retained a dense set of tokens for the *ACDC* sample, as shown in Fig. 5.7(b), and is highly concentrated within the target region. This helped preserve the finer details related to the different cardiac organs. In contrast, the framework could intelligently adapt to the homogeneous texture of the skin lesions and pruned redundant tokens related to the uniform structure of the ROI [Fig. 5.7(d)]. This illustrates that *PrATo* seamlessly adapts its pruning mechanism based on the characteristics of the target structures.

Fig. 5.8 shows the effectiveness of the scoring and weighting mechanism of *PrATo*. Fig. 5.8(a) illustrates that the token relevance ranking distribution is non-uniform, with a small subset of highly relevant tokens. The flat "True Uniform" line, marked in red, represents a vast majority of tokens having relevance scores close to zero. These scores might lead to instability while generating the pruning mask. Fig. 5.8(b) demonstrates the weight distribution generated by the proposed entropy-based weighting scheme. The transformed weights follow the true uniform line, which highlights that our *PrATo* could effectively transform the skewed weights to a nearly uniform distribution. Such a stable weighting system indicates a robust pruning strategy, as it prevents a few outlier tokens having extreme scores, from skewing the thresholding process.

The robustness analysis of the *PrATo* framework, under erroneous prompting conditions, is presented in Table 5.6. The performance of *PrATo* was analyzed across varying scenarios, *viz.*, tight, oversized, partial and misleading

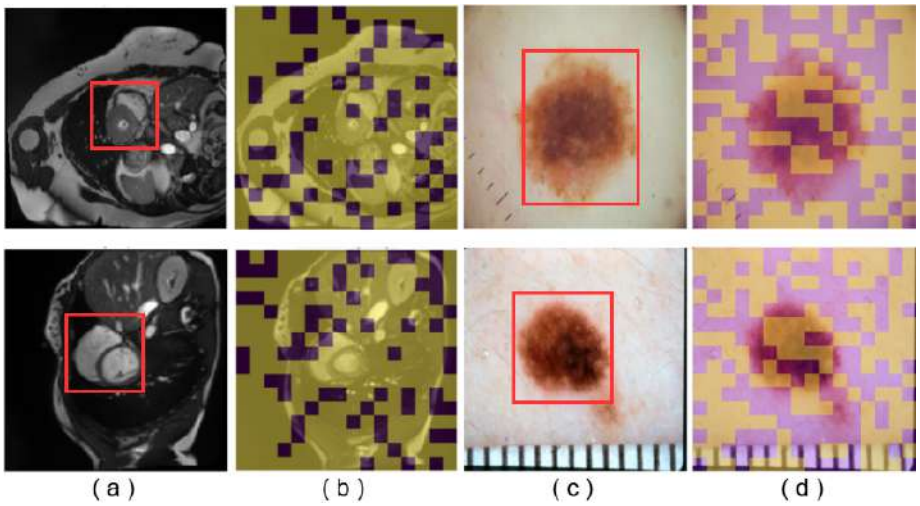

FIGURE 5.7: Sample token retention maps of *PrATo*, from (a)-(b) *ACDC* and (c)-(d) *ISIC* datasets. Red box highlights the target region in the input (a) and (c). Yellow patches in (b) and (d) represent the retained tokens by *PrATo*.

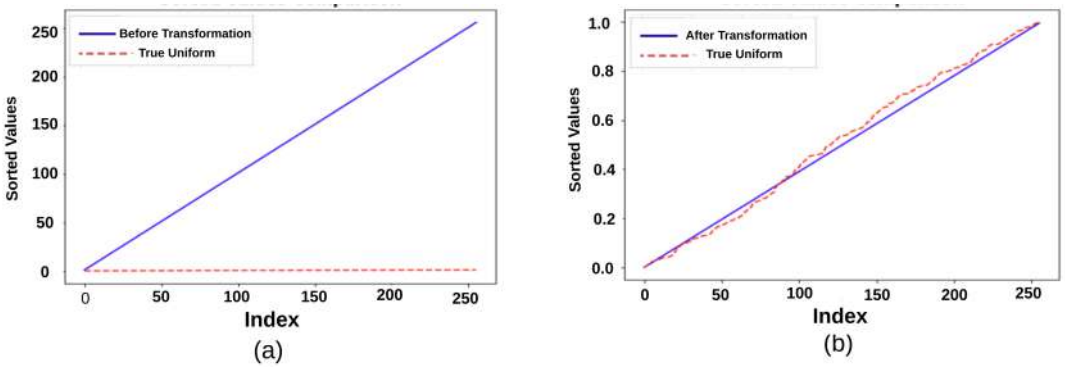

FIGURE 5.8: Line plots comparing the (a) non-uniform distribution of token relevance ranking, and (b) uniform distribution of entropy-based weighting, for an image.

(incorrectly placed box) prompts. The quantitative results suggest that *PrATo* exhibited graceful degradation when the box prompts were placed incorrectly. The partial prompts show significant robustness, with minimal drop in the mean values of DSC and IoU. This is because partial prompts still provide high-quality features related to a portion of the target structure. The global context, learned from the self-attention mechanism of ViT, helped to accurately predict the remaining regions. Oversized prompts include background information along with the target region, resulting in over-segmentation as evident in Fig. 5.9(a). The misleading prompts divert the attention of the framework away from the target region, resulting in the lowest average values of DSC and IoU. However, the segmentation model learns robust feature representations, strong enough to override the misleading spatial priors. This led to a reasonable performance. Although there are several segmentation errors [Fig. 5.9(d)], the overall structure is largely preserved. Thus, *PrATo* can handle imperfect prompts, common in interactive segmentation, without collapsing the overall

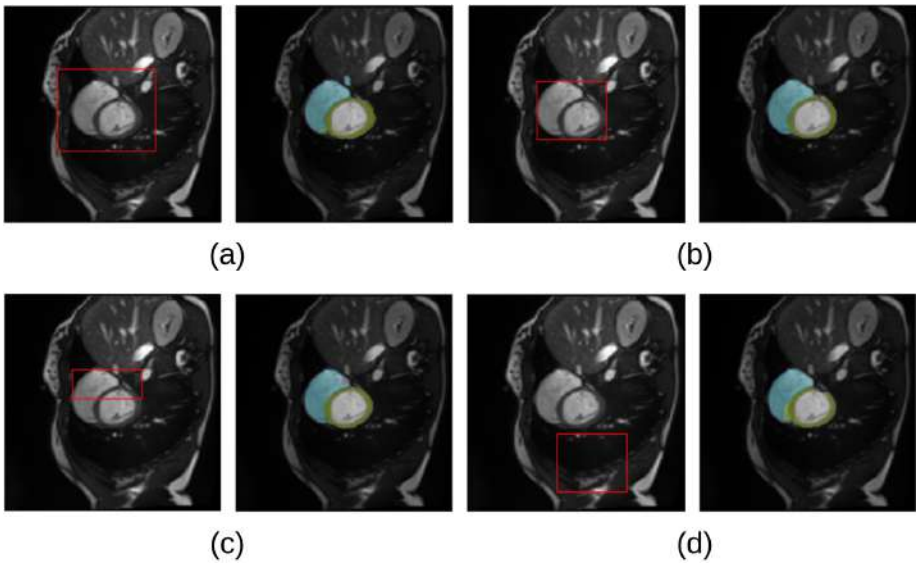

FIGURE 5.9: Qualitative results of *PrATo* under conditions of (a) oversized, (b) tight-boxed, (c) partial, and (d) misleading prompts.

TABLE 5.6: Quantitative analysis of *PrATo*, under different prompting conditions, on the *ACDC* dataset.

Approach	**mDSC**	**mIoU**	**mHD95**
Tight Prompts	**0.7169**	**0.6129**	19.87
Oversized Prompts	0.6571	0.552	21.43
Partial Prompts	0.675	0.5763	**17.16**
Misleading Prompts	0.6528	0.5489	19.31

performance of the underlying segmentation model.

5.4 Conclusion

This chapter presented *PrATo*, an adaptive framework for token pruning in Vision Transformers (ViT), designed to decrease the computational resources required to process irrelevant tokens. Using spatial priors through box prompts enhanced the retention of semantically relevant tokens, related to the target anatomical structure; thus, improving segmentation accuracy. The calculation of the entropy-guided similarity score, to identify tokens aligned with the spatial prior, ensured that the token selection process was data-driven. This allowed robustness to variations within the different input images. The experimental results suggested that the framework offers highly accurate segmentation performance and faster inference speed than other related approaches. Our strategy reduced computational costs, while maintaining (or even improving) the segmentation performance of several ViT-based segmentation models. This demonstrates the applicability of our *PrATo* framework across different architectures. *PrATo* shows robustness by demonstrating graceful degradation

with inaccurate prompts. Thus, this framework is highly relevant for automated medical image segmentation, as it can effectively circumvent the high computational overhead associated with ViTs. This makes them deployable in resource-constrained settings.

The performance of *PrATo* is highly dependent on the quality of the box-prompts. Furthermore, the dynamic hard pruning mechanism of *PrATo* may discard few fine-grained features related to the boundaries of anatomical structures. However, the framework can be improved by integrating fine-grained prompts, such as points and scribbles, to enhance the boundary delineation capacity while preserving efficient segmentation performance across a wide range of medical image analysis tasks and modalities.

Chapter 6

Conclusions and Future Scope

This final chapter summarizes the contributions of this thesis, evaluates the boundaries of the research and lays a clear path for extending this work in the future. The chapter begins with the *Conclusion* section, illustrating the key findings from previous chapters. Subsequently, the *Limitations* section outlines the broader boundaries of the developed models. These limitations lay the foundation for the *Future Scope* which presents a clear roadmap for developing the next generation of robust and clinically reliable segmentation models.

6.1 Conclusions

This thesis presented different algorithms for automated medical image segmentation, addressing the dual challenge of computational efficiency with the generation of high-quality feature representation. Different deep learning architectures were designed for efficient segmentation of anatomical structures having huge variability in their shapes and sizes. The models explored different architectural themes ranging from CNNs with a convolutional attention mechanism to integrating spatial-spectral features without excessive cost. The study then investigated hybridizing next-generation sequential models, such as Vision-xLSTM with CNNs, to develop computationally efficient segmentation methods. Algorithmic optimization techniques, such as prompt-driven token pruning methods, were introduced to reduce the computational load of state-of-the-art ViT-based segmentation methods. This section summarizes the key findings from these studies, and demonstrates the efficiency of the developed models across different datasets. It highlights their collective contribution towards deployable computer-aided systems.

- The initial study was in developing a novel global-context aware convolutional attention mechanisms to achieve high segmentation accuracy across 2D and 3D datasets. The research began with the development of Full Scale Deeply Supervised Attention Network (*FuDSA-Net*), a 2D framework to delineate complex COVID-19 and minute DR lesions from lung CT and eye fundus images, respectively. The core contribution was the development of a novel convolutional attention mechanism, which refined the feature responses from different levels of the low-projection path (encoder). The "full scale" feature gamut led to the integration of coarser global details and finer local information for a robust feature representation. This approach, along with deep supervision and residual connections in the high-projection path (decoder), proved superior in delineating complex lesion boundaries compared to state-of-the-art architectures.

  The framework was then extended to the integrated global-Context volumetric Attention Network (*VoCANet*) for volumetric image segmentation. It also made use of the multi-scalar attention module to segment abdominal organs and adrenal gland tumors of variable texture and geometry. The model was specifically designed to optimize computational resources needed for processing volumetric data, resulting in lower TFLOPs and parameter count compared to baselines.

- Next, the research addressed the limitations of segmentation models that analyze only the spatial domain information of the input image. The Wavelet-infused Convolutional Transformer (*WaveCoformer*) was developed. A novel module, Spectral Feature Convolution (*SpectraConv*), was designed to process the high-frequency (textural information) and low-frequency components (structural information) derived from the Discrete Wavelet Transform (DWT) to learn fine-grained textural patterns of the target regions. In parallel, the Dual Attention (*DA*) block identified relevant channels from the feature gamut followed by the modeling of global structural information from the spatial domain using Swin Transformers. This helped in reducing the quadratic computational complexity of the self-attention mechanism, by using a window-based attention computation approach. Finally, the bi-directional fusion module Cross Context Attention (*CCA*), efficiently combined the spectral and spatial domain features from *SpectraConv* and *DA* module to generate robust representations. The model achieved superior segmentation accuracy, along with the lowest parameter count and TFLOPs, compared to other baseline methods. The computational efficiency was attributed to the use of depthwise separable and dilated convolution kernels, along with the window-based self-attention computation approach.

- The study then shifted towards exploring computationally efficient alternatives to ViTs, by focusing on Vision Extended Long Short Term Memory (Vision-xLSTM). This led to the development of *U*-Vision-xLSTM (*U-VixLSTM*), a pioneering study towards integrating CNN with Vision-xLSTM for medical image segmentation. The CNN-based encoder path of the $U-$shaped framework captured fine-grained patterns in the generated feature maps. Next, the Vision-xLSTM blocks modeled the global dependency within these activation maps. The linear computational and constant memory complexity of xLSTM overcame the quadratic analytical complexity of the self-attention mechanism. Building on this approach, the Vision-xLSTM blocks were advanced to learn cross-dimensional dependencies in volumetric inputs through the development of the Rotational $U-$Vision-xLSTM (*Rot-UViL*). The novel Rotational Attention Module (*RAM*) had three different pathways of Vision-xLSTM blocks. While one path modeled the standard spatial dependency between height and width axes, the other two pathways rotated the tensor to learn the dependencies between channel and height/width axes; thereby, developing a holistic understanding of the volumetric structures spanning across the entire input volume. Both approaches strongly aligned with the central theme of the thesis; *viz.* achieving superior segmentation accuracy while being computationally efficient, with lower parameters, FLOPs, and memory footprint, as compared to their state-of-the-art ViT counterparts.

- Finally, the thesis addressed the challenge of computational efficiency by developing an algorithmic framework to optimize the state-of-the-art ViT-based segmentation frameworks. The Prompt-driven Adaptive Token

pruning method (*PrATo*) was designed to dynamically reduce the processing of semantically irrelevant tokens across the segmentation pipeline. The algorithm adopted a parameter-free scoring mechanism to rank the tokens based on their relevance. The box prompt acted as a spatial prior to identify the relevant region in the entire input image. The process computed a similarity map between the prompted region features and the ViT-generated tokens. This was followed by an entropy-based scoring mechanism to identify tokens with high association to the prompted regions, which were to be retained in the segmentation pipeline. This data-driven strategy led to a reduction of 35-55% of processed tokens across different state-of-the-art segmentation methods. It significantly lowered computational costs while maintaining segmentation accuracy. This validated the generalizability of *PrATo* in making models efficient for deployment in resource-constrained settings.

The architectures developed in this thesis progressed through distinct phases of feature representation. The learning in *VoCANet* was realized by aggregating multi-scalar features for combining coarse global details and fine-grained information for robust representation, without relying on heavy models. As research progressed towards hybrid models, a contrast between key paradigms in medical image segmentation emerged. CNNs are proficient at extracting localized, fine-grained textural features but fail to model long-range anatomical dependencies. On the other hand, ViTs effectively capture global context through self-attention, yet their quadratic computational complexity results in significant computational and memory constraints. The advanced models in this study, namely *WaveCoformer* and *Rot-UViL*, learn by addressing this gap. *WaveCoformer* removes the computational bottleneck by substituting global attention with a highly efficient window-based computational method integrated with spatial-spectral filters. Ultimately, *Rot-UViL* employs the Vision-xLSTM framework, which substitutes conventional self-attention with exponential gating and matrix memory architectures. This mechanism enables the network to process sequences with linear computational complexity.

The frameworks discussed in this thesis are well-suited for real-world deployment due to their high segmentation accuracy along with computational efficiency. By introducing novel approaches such as multi-scalar convolutional attention, spectral-spatial fusion, cross-dimensional rotational attention, and efficient Vision-xLSTM blocks, the research has introduced novel tools for efficient 2D/3D input segmentation. The high segmentation accuracy is suitable for timely diagnosis and improved surgical planning. The lightweight automated approaches would assist medical practitioners by reducing the time and labor of manual segmentation. It facilitates accessibility of advanced healthcare tools, by improving patient care with optimized utilization of computing resources.

6.2 Limitations

While this thesis has presented a series of novel computationally efficient segmentation methods, there exist specific limitations as in any focused research

study. Acknowledging these boundaries is essential to lay the foundation for the future research directions, described in the subsequent section. The following limitations are based on the broader scope of the thesis.

- **Generalizability and domain shift:** The models developed in this thesis were trained and validated on high-quality and well-annotated datasets like *Synapse*, *AdrenalSeg*, etc. However, the methods lack large-scale pretraining, which makes them susceptible to domain-shift oriented problems, prevalent in real-world clinical data.
- **Lack of uncertainty quantification:** The developed methods lack the evaluation of their own uncertainty or confidence in the predicted segmentation mask. In real-world healthcare settings, determining the locations where the model is uncertain is equally important as the prediction. This helps identify the ambiguous cases, which need expert review.
- **Limited prompts as spatial prior:** The *PrATo* framework used only a box prompt as the spatial prior. This may be less precise than the other prompting approaches. The study did not include other fine-grained prompts like points or scribbles.
- **Dependence on supervised learning:** The models in this research were trained and validated using a fully supervised learning paradigm. This approach requires large, well-annotated datasets, and are expensive and time-consuming to create in the medical domain.
- **Interpretability challenges:** Analogous to many sequence-modeling architectures, the internal matrix memory states of Vision-xLSTM are intrinsically less interpretable than conventional CNN feature maps. In clinical environments, the black-box characteristic of these continuous memory updates poses an obstacle to direct clinical trust.

6.3 Future Scope

Although the contributions of this thesis are robust, they lead to several promising future directions. The limitations listed in the previous section provide a direct path for extending this research. The following points outline the path of developing efficient segmentation models for clinical deployment.

- **Convergence with Foundation Models:** The recent increasing popularity of medical vision foundation models requires a strategic deployment approach based on modality-to-task ratio. These large models are vital for multi-modality/anatomy and generalization across diverse tasks [116]. Fine-tuned versions [120] are preferable for single modality and multi-task scenarios. However, they remain computationally heavy for single-modality/anatomy, single-task workflows. For instance, in time-critical scenarios, like Image-Guided Therapy (IGT), real-time anatomical tracking is needed where deploying parameter-heavy models becomes

impractical. The resource-efficient models developed in this thesis become the optimal choice for such use cases. Future work will include bridging these paradigms via knowledge distillation, with heavy-foundation models as teachers for transferring their generalized representations into our lightweight student networks.

- **Enhancing domain generalization:** A critical step will be to enhance the out-of-distribution reliability, by integrating state-of-the-art domain generalization techniques into the novel architectures. This will ensure generalizable performance across large-scale, multi-center datasets.
- **Exploring self-supervised paradigms:** The study will be extended to incorporate self-supervised learning paradigms, to address the dependence on fully-annotated datasets.
- **Improving reliability of the models:** The deterministic networks, discussed in this thesis, can result in confident but incorrect predictions. Research can explore uncertainty-aware knowledge distillation by leveraging a large-scale pretrained teacher (like foundation models) to generate uncertainty maps along with the predictions. The lightweight models, developed in this thesis, can learn to match the segmentation accuracy of the teacher model, along with its confidence, for a clinically reliable tool.
- **Pruning with semantic relevance:** The extension of the pruning framework can be based entirely on semantic relevance. This involves learning human-understandable concepts to identify the relevant components of the model. These high-level concepts will make the pruning process transparent – an essential requirement in the medical domain for unbiased decision-making.
- **Interactive segmentation workflows:** Another important extension is to create a human-in-the-loop segmentation framework. The domain expert can provide corrective prompts, in the event of an erroneous prediction. The model would then incorporate the feedback to refine the segmentation mask.
- **Leveraging complementary features with ensembling:** Each of the developed models in this thesis possesses its own unique strengths, ranging from learning multi-scalar features to modeling global context with linear complexity. A promising direction can be exploring an ensemble to aggregate the predictions from these diverse backbones. This would lead to a robust and accurate outcome, when compared to the performance of individual models.
- **Hardware-aware implementation:** A critical research direction could be to transition from pure software-level architecture design to hardware-aware implementations, like neuromorphic computing [32], suitable for portable medical devices.

List of Publications

1. Published/Accepted:

Journals:

J1. P. Dutta, S. Mitra, and S. K. Roy, "Wavelet-infused convolution- transformer for efficient segmentation in medical images," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 55, pp. 3326–3337, 2025. doi:10.1109/TSMC.2025.3539573.

J2. P. Dutta, S. Bose, S. K. Roy, and S. Mitra, "Are vision-xLSTM-embedded U- Nets better at segmenting medical images?," *Neural Networks*, p. 107925, 2025.doi:10.1016/j.neunet.2025.107925.

Conferences:

C1. S. Dey, P. Dutta, S. Mitra, and B. U. Shankar, "Multi-scale deep supervised attention network for red lesion segmentation," in *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1–4, 2023. doi:10.1109/ISBI53787.2023.10230639.

C2. P. Dutta and S. Mitra, "Full-scale deeply supervised attention network for segmenting COVID- 19 lesions," in *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1–4, 2023. doi:10.1109/ISBI53787.2023.10230579.

C3. P. Dutta and S. Mitra, "Efficient global-context driven volumetric segmentation of abdominal images," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1880–1885, 2023. doi:10.1109/BIBM58861.2023.10385802.

C4. P. Dutta and S. Mitra, "Exploiting cross-dimensional dependency using vision-LSTM for efficient medical image segmentation," in *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1–4, IEEE, 2025. doi:10.1109/ISBI60581.2025.10981187.

C5. S. Dey, P. Dutta, R. Bhattacharyya, S. Pal, S. Mitra, R. Raman, " Adaptive Class Learning to Screen Diabetic Disorders in Fundus Images of Eye, " in *Proceedings of International Conference on Pattern Recognition (ICPR)*, 2024, Springer, doi:10.1007/978-3-031-78104-9_9.

C6. A. Maity, P. Dutta and S. Mitra, "EcoMamba-Net: A Parameter- Efficient Architecture for Medical Image Segmentation," in *Proceedings of International Conference on Pattern Recognition and Machine Intelligence*, 2025. *(accepted)*

2. Under Processing:

Journals:

J1. P. Dutta, A. Maity, and S. Mitra, "Prompt-based dynamic token pruning for efficient segmentation of medical images," *arXiv preprint arXiv:2506.16369*, 2025. doi:10.48550/arXiv.2506.16369. *(under review)*

Patents Filed:

P1. S. Mitra and P. Dutta, " System and Method for Effective Volumetric Segmentation of Medical Images involving Spatio-Spectral Components", Indian Patent Application # 202431005490 dt. January 27, 2024.

Bibliography

[1] N. Abraham and N. M. Khan, "A novel Focal Tversky loss function with improved Attention *U*-Net for lesion segmentation," in *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, 2019, pp. 683–687.

[2] R. Adams and L. Bischof, "Seeded region growing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, pp. 641–647, 1994.

[3] B. Alkin, M. Beck, and *et al.*, "Vision-LSTM: XLSTM as generic vision backbone," in *Proceedings of the International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=SiH7DwNKZZ.

[4] S. Aminizadeh, A. Heidari, and *et al.*, "Opportunities and challenges of artificial intelligence and distributed systems to improve the quality of healthcare service," *Artificial Intelligence in Medicine*, vol. 149, p. 102 779, 2024.

[5] A. Angelopoulou, A. Psarrou, and *et al.*, "3D reconstruction of medical images from slices automatically landmarked with growing neural models," *Neurocomputing*, vol. 150, pp. 16–25, 2015.

[6] A. Arbelle, S. Cohen, and *et al.*, "Dual-task ConvLSTM-UNet for instance segmentation of weakly annotated microscopy videos," *IEEE Transactions on Medical Imaging*, vol. 41, pp. 1948–1960, 2022.

[7] R. Azad, E. K. Aghdam, and *et al.*, "Medical image segmentation review: The success of *U-N*et," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, pp. 10 076–10 095, 2024.

[8] J. L. Ba, J. R. Kiros, and *et al.*, "Layer Normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[9] R. Bagaria, S. Wadhwani, and *et al.*, "A wavelet-based segmentation technique for medical images," in *Proceedings of the Artificial Intelligence and Sustainable Computing*, Springer, 2021, pp. 65–77.

[10] S. Banerjee, S. Mitra, and *et al.*, "Automated 3D segmentation of brain tumor using visual saliency," *Information Sciences*, vol. 424, pp. 337–353, 2018.

[11] M. Beck, K. Pöppel, and *et al.*, "XLSTM: Extended Long Short-Term Memory," in *Proceedings of the Conference on Neural Information Processing Systems*, 2024. [Online]. Available: https://openreview.net/pdf?id=ARAxPPIAhq.

[12] O. Bernard, A. Lalande, and *et al.*, "Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved?" *IEEE Transactions on Medical Imaging*, vol. 37, pp. 2514–2525, 2018. DOI: 10.1109/TMI.2018.2837502.

[13] J. C. Bezdek, L. Hall, and *et al.*, "Review of MR Image segmentation techniques using pattern recognition.," *Medical Physics*, vol. 20, pp. 1033–1048, 1993.

[14] Y. Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in ND images," in *Proceedings of the Eighth IEEE International Conference on Computer Vision ICCV*, IEEE, vol. 1, 2001, pp. 105–112.

[15] Y. Cai and Y. Wang, "MA-UNet: An improved version of *U*-Net based on multi-scale and attention mechanism for medical image segmentation," in *Proceedings of the International Conference on Electronics and Communication; Network and Computer Technology (ECNCT)*, SPIE, vol. 12167, 2022, pp. 205–211.

[16] G. Carlos, K. Figueiredo, A. Hussain, and M. Vellasco, "SegQNAS: Quantum-inspired Neural Architecture Search applied to medical image semantic segmentation," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 2023, pp. 1–8.

[17] S. N. Chandrasekaran, H. Ceulemans, and *et al.*, "Image-based profiling for drug discovery: Due for a machine-learning upgrade?" *Nature Reviews Drug Discovery*, vol. 20, pp. 145–159, 2021.

[18] B. Chen, Y. Liu, and *et al.*, "TransAttUNet: Multi-level attention-guided *U*-Net with Transformer for medical image segmentation," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 8, pp. 55–68, 2023.

[19] J. Chen, Y. Lu, and *et al.*, "TransUNet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[20] J. Chen, X. Zhang, and *et al.*, "TA-ASF: Attention-sensitive token sampling and fusing for visual Transformer models on the edge," in *Proceedings of the IEEE/ACM Symposium on Edge Computing (SEC)*, IEEE, 2024, pp. 123–134.

[21] Y. Chen, T. Zhou, and *et al.*, "HADCNet: Automatic segmentation of COVID-19 infection based on a hybrid attention dense connected network with dilated convolution," *Computers in Biology and Medicine*, vol. 149, p. 105 981, 2022.

[22] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1251–1258.

[23] F. Chollet, *Deep Learning with Python*. New York, NY: Manning, 2021.

[24] D.-A. Clevert, T. Unterthiner, and *et al.*, "Fast and accurate deep network learning by exponential linear units (ELUs)," *arXiv preprint arXiv:1511.07289*, 2015.

[25] N. C. Codella, D. Gutman, and *et al.*, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), hosted by the International Skin Imaging Collaboration (ISIC)," in *Proceedings of the IEEE 15th International Symposium on Biomedical Imaging (ISBI)*, IEEE, 2018, pp. 168–172. DOI: 10.1109/ISBI.2018.8363547.

[26] J. Dai, T. Liu, D. A. Torigian, Y. Tong, S. Han, P. Nie, J. Zhang, R. Li, F. Xie, and J. K. Udupa, "GA-Net: A geographical attention neural network for the segmentation of body torso tissue composition," *Medical Image Analysis*, vol. 91, p. 102 987, 2024.

[27] T. Das and P. Guha, "The puzzle of public health expenditure and healthcare infrastructure in India: An empirical investigation," *Regional Science Policy & Practice*, vol. 16, p. 12 710, 2024.

[28] E. Decencière, X. Zhang, and *et al.*, "Feedback on a publicly distributed image database: The Messidor database," *Image Analysis & Stereology*, vol. 33, pp. 231–234, 2014.

[29] J. Deng, W. Dong, and *et al.*, "ImageNet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2009, pp. 248–255.

[30] S. Dey, P. Dutta, and *et al.*, "Multi-scale deep supervised attention network for red lesion segmentation," in *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, 2023, pp. 1–4. DOI: 10.1109/ISBI53787.2023.10230639.

[31] P. Dhar, "The carbon impact of artificial intelligence," *Nature Machine Intelligence*, pp. 423–425, 2020.

[32] B. Ding, L. Chen, C. Li, T. Huang, and *et al.*, "Memristor-based selective convolutional circuit for high-density salt-and-pepper noise removal," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 72, pp. 3115–3125, 2025. DOI: 10.1109/TCSI.2025.3566364.

[33] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[34] A. Dosovitskiy, L. Beyer, and *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy.

[35] P. Dutta, S. Bose, and *et al.*, "Are vision-xLSTM-embedded *U*-Nets better at segmenting medical images?" *Neural Networks*, p. 107 925, 2025.

[36] P. Dutta, A. Maity, and *et al.*, "Prompt-based dynamic token pruning for efficient segmentation of medical images," *arXiv preprint arXiv:2506.16369*, 2025.

[37] P. Dutta and S. Mitra, "Efficient global-context driven volumetric segmentation of abdominal images," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2023, pp. 1880–1885. DOI: 10.1109/BIBM58861.2023.10385802.

[38] P. Dutta and S. Mitra, "Exploiting cross-dimensional dependency using vision-LSTM for efficient medical image segmentation," in *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, IEEE, 2025, pp. 1–4.

[39] P. Dutta and S. Mitra, "Full-scale deeply supervised attention network for segmenting COVID-19 lesions," in *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, 2023, pp. 1–4. DOI: 10.1109/ISBI53787.2023.10230579.

[40] P. Dutta, S. Mitra, and *et al.*, "Wavelet-infused convolution-transformer for efficient segmentation in medical images," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 55, pp. 3326–3337, 2025. DOI: 10.1109/TSMC.2025.3539573.

[41] S. Elfwing, E. Uchibe, and *et al.*, "Sigmoid-weighted linear units for neural network function approximation in reinforcement learning," *Neural Networks*, vol. 107, pp. 3–11, 2018. DOI: https://doi.org/10.1016/j.neunet.2017.12.012.

[42] J. L. Elman, "Distributed representations, simple recurrent networks, and grammatical structure," *Machine Learning*, vol. 7, pp. 195–225, 1991.

[43] J. Fan, C. Li, and *et al.*, "ScaleKD: Strong vision transformers could be excellent teachers," in *Proceedings of the Thirty-Eighth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024. [Online]. Available: https://openreview.net/forum?id=0WCFI2Qx85.

[44] A. F. Frangi, W. J. Niessen, and *et al.*, "Multiscale vessel enhancement filtering," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Springer, 1998, pp. 130–137.

[45] K. Fujihara, Y. Matsubayashi, M. H. Yamada, M. Yamamoto, T. Iizuka, K. Miyamura, Y. Hasegawa, H. Maegawa, S. Kodama, T. Yamazaki, and *et al.*, "Machine learning approach to decision making for insulin initiation in Japanese patients with type 2 diabetes (JDDM 58): Model development and validation study," *JMIR Medical Informatics*, vol. 9, e22148, 2021.

[46] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*. New York, NY: Prentice-Hall, Inc., 2006.

[47] I. Goodfellow, Y. Bengio, and *et al.*, *Deep Learning*. MIT Press Cambridge, 2016, vol. 1.

[48] O. Gozes, M. Frid-Adar, and *et al.*, "Rapid AI development cycle for the Coronavirus (COVID-19) pandemic: Initial results for automated detection & patient monitoring using deep learning CT image analysis," *arXiv:2003.05037*, 2020. [Online]. Available: https://spectrum.ieee.org/hospitals-deploy-ai-tools-detect-covid19-chest-scans.

[49] S. Grudzenski, M. A. Kuefner, and *et al.*, "Contrast medium–enhanced radiation damage caused by CT examinations," *Radiology*, vol. 253, pp. 706–714, 2009.

[50] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," in *Proceedings of the Conference on Language Modeling (COLM)*, 2024. [Online]. Available: https://openreview.net/forum?id=tEYskw1VY2.

[51] Z. Gu, J. Cheng, and *et al.*, "CE-Net: Context Encoder Network for 2D medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 38, pp. 2281–2292, 2019. DOI: 10.1109/TMI.2019.2903562.

[52] A. Hatamizadeh, V. Nath, and *et al.*, "Swin UNETR: Swin Transformers for semantic segmentation of brain tumors in MRI images," in *Proceedings of the International MICCAI Brainlesion Workshop*, Springer, 2021, pp. 272–284.

[53] A. Hatamizadeh, Y. Tang, and *et al.*, "UNETR: Transformers for 3D medical image segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, (WACV)*, 2022, pp. 574–584.

[54] K. He, G. Gkioxari, and *et al.*, "Mask R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2961–2969.

[55] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," *arXiv preprint arXiv:1606.08415*, 2016.

[56] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, pp. 1735–1780, 1997.

[57] K. Hornik, M. Stinchcombe, and *et al.*, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, pp. 359–366, 1989.

[58] H. Hu, C. Yu, and *et al.*, "HDConv: Heterogeneous kernel-based dilated convolutions," *Neural Networks*, vol. 179, p. 106 568, 2024.

[59] J. Hu, L. Shen, and *et al.*, "Squeeze-and-Excitation networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141. DOI: 10.1109/CVPR.2018.00745.

[60] H. Huang, L. Lin, and *et al.*, "UNet 3+: A full-scale connected *U*-Net for medical image segmentation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 1055–1059.

[61] W. Huang, Y. Shen, and *et al.*, "A general and efficient training for Transformer via token expansion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 15 783–15 792.

[62] S. Iqbal, T. M. Khan, and *et al.*, "TBConvL-Net: A hybrid deep learning architecture for robust medical image segmentation," *Pattern Recognition*, vol. 158, p. 111 028, 2025.

[63] M. Jia, L. Tang, and *et al.*, "Visual prompt tuning," in *Proceedings of the European Conference on Computer Vision*, Springer, 2022, pp. 709–727.

[64] X. Jiang, J. Jiang, and *et al.*, "SEACU-Net: Attentive ConvLSTM *U*-Net with squeeze-and-excitation layer for skin lesion segmentation," *Computer Methods and Programs in Biomedicine*, vol. 225, p. 107 076, 2022.

[65] L. Kang, Z. Zhou, and *et al.*, "Renal tumors segmentation in abdomen CT images using 3D-CNN and ConvLSTM," *Biomedical Signal Processing and Control*, vol. 72, p. 103 334, 2022.

[66] M. Kass, A. Witkin, and *et al.*, "Snakes: Active contour models," *International Journal of Computer Vision*, vol. 1, pp. 321–331, 1988.

[67] A. Krizhevsky, I. Sutskever, and *et al.*, "ImageNet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems (NIPS)*, vol. 25, 2012.

[68] B. Landman, Z. Xu, and *et al.*, "MICCAI multi-atlas labeling beyond the cranial vault–workshop and challenge," in *Proceedings of MICCAI Multi-Atlas Labeling beyond Cranial Vault—Workshop Challenge*, 2015, p. 12.

[69] Y. LeCun, Y. Bengio, and *et al.*, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[70] Y. LeCun, L. Bottou, and *et al.*, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, pp. 2278–2324, 1998.

[71] J. Lei Ba, J. R. Kiros, and *et al.*, "Layer normalization," *ArXiv e-prints*, arXiv–1607, 2016.

[72] W. Li, S. Wen, K. Shi, Y. Yang, and T. Huang, "Neural Architecture Search with a lightweight Transformer for text-to-image synthesis," *IEEE Transactions on Network Science and Engineering*, vol. 9, pp. 1567–1576, 2022. DOI: 10.1109/TNSE.2022.3147787.

[73] A. Lin, B. Chen, and *et al.*, "DS-TransUNet: Dual Swin Transformer *U*-Net for medical image segmentation," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–15, 2022.

[74] J. Liu, H. Yang, and *et al.*, "Swin UMamba: Mamba-based *U*-Net with ImageNet-based pretraining," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Springer, 2024, pp. 615–625.

[75] T. Liu, Q. Bai, D. A. Torigian, Y. Tong, and J. K. Udupa, "VSmTrans: A hybrid paradigm integrating self-attention and convolution for 3D medical image segmentation," *Medical Image Analysis*, vol. 98, p. 103 295, 2024.

[76] Z. Liu, Y. Lin, and *et al.*, "Swin Transformer: Hierarchical Vision Transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 10 012–10 022.

[77] J. Ma, C. Ge, and *et al.*, *COVID-19 CT Lung and Infection Segmentation Dataset*, https://doi.org/10.5281/zenodo.3757476, 2020. DOI: 10.5281/zenodo.3757476. [Online]. Available: https://doi.org/10.5281/zenodo.3757476.

[78] J. Ma, F. Li, and *et al.*, "*U*-Mamba: Enhancing long-range dependency for biomedical image segmentation," *arXiv preprint arXiv:2401.04722*, 2024.

[79] A. L. Maas, A. Y. Hannun, and *et al.*, "Rectifier nonlinearities improve neural network acoustic models," in *Proceedings of the 30th International Conference on Machine Learning (ICML)*, Atlanta, GA, 2013, p. 3.

[80] S. G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, pp. 674–693, 1989.

[81] M. Marchetti, D. Traini, and *et al.*, "Efficient token pruning in Vision Transformers using an attention-based multilayer network," *Expert Systems with Applications*, vol. 279, p. 127 449, 2025.

[82] MedSeg, H. B. Jenssen, and *et al.*, *MedSeg COVID Dataset 1*, https://figshare.com/articles/dataset/MedSeg_Covid_Dataset_1/13521488, 2021. DOI: 10.6084/m9.figshare.13521488.v2.

[83] MedSeg, H. B. Jenssen, and *et al.*, *MedSeg COVID Dataset 2*, https://figshare.com/articles/dataset/Covid_Dataset_2/13521509, 2021. DOI: 10.6084/m9.figshare.13521509.v2.

[84] W. Merrill, J. Petty, and *et al.*, "The illusion of state in State-Space Models," in *Proceedings of the International Conference on Machine Learning (ICML)*, PMLR, 2024, pp. 35 492–35 506. DOI: https://doi.org/10.48550/arXiv.2404.08819.

[85] F. Milletari, N. Navab, and *et al.*, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proceedings of the International Conference on 3D Vision (3DV)*, IEEE, 2016, pp. 565–571.

[86] M. Minsky and S. Papert, *Perceptrons: An Introduction to Computational Geometry*. Cambridge, MA, USA: MIT Press, 1969.

[87] D. Mishra, S. Chaudhury, and *et al.*, "Segmentation of vascular regions in ultrasound images: A deep learning approach," in *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS)*, IEEE, 2018, pp. 1–5.

[88] D. Mishra, S. Chaudhury, and *et al.*, "Ultrasound image segmentation: A deeply supervised network with attention to boundaries," *IEEE Transactions on Biomedical Engineering*, vol. 66, pp. 1637–1648, 2018.

[89] T. M. Mitchell, *Machine Learning*. New York, NY: McGraw-Hill, 1997.

[90] S. Mitra and B. U. Shankar, "Integrating radio imaging with gene expressions toward a personalized management of cancer," *IEEE Transactions on Human-Machine Systems*, vol. 44, pp. 664–677, 2014. DOI: 10.1109/THMS.2014.2325744.

[91] A. W. Moawad, A. A. Ahmed, and *et al.*, "Voxel-level segmentation of pathologically-proven Adrenocortical carcinoma with Ki-67 expression (Adrenal-ACC-Ki67-Seg) [Data set]," *The Cancer Imaging Archive*, 2023. [Online]. Available: https://doi.org/10.7937/1FPG-VM46.

[92] S. P. Morozov, A. E. Andreychenko, and *et al.*, "MOSMED data: Data set of 1110 chest CT scans performed during the COVID-19 epidemic," *Digital Diagnostics*, vol. 1, pp. 49–59, 2020.

[93] J. Mukherjee and S. K. Mitra, "Enhancement of color images by scaling the DCT coefficients," *IEEE Transactions on Image Processing*, vol. 17, pp. 1783–1794, 2008.

[94] S. A. Nadeem, E. A. Hoffman, J. C. Sieren, A. P. Comellas, S. P. Bhatt, I. Z. Barjaktarevic, F. Abtin, and P. K. Saha, "A CT-based automated algorithm for airway segmentation using freeze-and-grow propagation and deep learning," *IEEE Transactions on Medical Imaging*, vol. 40, pp. 405–418, 2020.

[95] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2010, pp. 807–814.

[96] S. Niyas, S. Pawan, and *et al.*, "Medical image segmentation with 3D convolutional neural networks: A survey," *Neurocomputing*, vol. 493, pp. 397–413, 2022.

[97] O. Oktay, J. Schlemper, and *et al.*, "Attention U-Net: Learning where to look for the pancreas," in *Proceedings of the Medical Imaging with Deep Learning (MIDL)*, 2018. DOI: https://doi.org/10.48550/arXiv.1804.03999.

[98] N. Park and S. Kim, "How do Vision Transformers work?" In *Proceedings of International Conference on Learning Representations*, 2021. [Online]. Available: https://arxiv.org/abs/2202.06709.

[99] C. Parmar, E. Rios Velazquez, and *et al.*, "Robust radiomics feature quantification using semiautomatic volumetric segmentation," *PLOS One*, vol. 9, e102107, 2014.

[100] J. Pichat, J. E. Iglesias, and *et al.*, "A survey of methods for 3D histology reconstruction," *Medical Image Analysis*, vol. 46, pp. 73–105, 2018.

[101] P. Porwal, S. Pachade, and *et al.*, "Indian Diabetic Retinopathy Image Dataset (IDRID): A database for Diabetic Retinopathy screening research," *Data*, vol. 3, p. 25, 2018.

[102] Y. Qiu, Y. Liu, and *et al.*, "Miniseg: An extremely minimum network for efficient COVID-19 segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 4846–4854.

[103] Y. Rao, W. Zhao, and *et al.*, "DynamicViT: Efficient vision transformers with dynamic token sparsification," in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2021. [Online]. Available: https://openreview.net/forum?id=jB0Nlbwlybm.

[104] S. Ren, D. Zhou, and *et al.*, "Shunted self-attention via multi-scale token aggregation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10 853–10 862.

[105] E. Rich, K. Knight, and *et al.*, *Artificial Intelligence*. New Delhi: Tata McGraw-Hill Education, 2009.

[106] O. Ronneberger, P. Fischer, and *et al.*, "*U*-Net: Convolutional networks for biomedical image segmentation," in *Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI*, Springer, 2015, pp. 234–241.

[107] J. Rösch, F. S. Keller, and J. A. Kaufman, "The birth, early years, and future of interventional radiology," *Journal of Vascular and Interventional Radiology*, vol. 14, pp. 841–853, 2003.

[108] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," *Psychological Review*, vol. 65, pp. 386–408, 1958.

[109] A. G. Roy, N. Navab, and *et al.*, "Concurrent spatial and channel 'squeeze & excitation' in fully convolutional networks," in *Proceedings of the Medical Image Computing and Computer Assisted Intervention (MICCAI)*, Springer, 2018, pp. 421–429.

[110] D. E. Rumelhart, G. E. Hinton, and *et al.*, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, 1986.

[111] P. K. Saha, S. A. Nadeem, and *et al.*, "A survey on artificial intelligence in pulmonary imaging," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 13, e1510, 2023.

[112] T. N. Sainath, O. Vinyals, and *et al.*, "Convolutional, long short-term memory, fully connected deep neural networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2015, pp. 4580–4584.

[113] S. S. M. Salehi, D. Erdogmus, and *et al.*, "Tversky loss function for image segmentation using 3D fully convolutional deep networks," in *Proceedings of the International Workshop on Machine Learning in Medical Imaging*, Springer, 2017, pp. 379–387.

[114] N. Salpea, P. Tzouveli, and *et al.*, "Medical image segmentation: A review of modern architectures," in *European Conference on Computer Vision*, Springer, 2022, pp. 691–708.

[115] G. Satya Nugraha, B. Amelia Riyandari, and *et al.*, "RGB channel analysis for glaucoma detection in retinal fundus image," in *Proceedings of the International Conference on Advancement in Data Science, E-learning and Information Systems (ICADEIS)*, 2020, pp. 1–5. DOI: 10.1109/ICADEIS49811.2020.9277230.

[116] A. Sellergren, S. Kazemzadeh, and *et al.*, "MedGemma technical report," *arXiv preprint arXiv:2507.05201*, 2025.

[117] A. Shaker, M. Maaz, and *et al.*, "UNETR++: Delving into efficient and accurate 3D medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 43, pp. 3377–3390, 2024.

[118] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, 1948.

[119] E. Shelhamer, J. Long, and *et al.*, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 640–651, 2017. DOI: 10.1109/TPAMI.2016.2572683.

[120] H. Shu, W. Li, and *et al.*, "TinySAM: Pushing the envelope for efficient segment anything model," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, 2025, pp. 20 470–20 478.

[121] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of the International Conference on Learning Representations ICLR*, 2015. [Online]. Available: http://arxiv.org/abs/1409.1556.

[122] P. Suetens, *Fundamentals of Medical Imaging*. Cambridge: Cambridge University Press, 2017.

[123] C. Szegedy, W. Liu, and *et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9. DOI: 10.1109/CVPR.2015.7298594.

[124] S. A. Taghanaki, Y. Zheng, and *et al.*, "Combo loss: Handling input and output imbalance in multi-organ segmentation," *Computerized Medical Imaging and Graphics*, vol. 75, pp. 24–33, 2019.

[125] Q. Tang, B. Zhang, and *et al.*, "Dynamic token pruning in plain vision transformers for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 777–786.

[126] A. R. Teymurazyan, R. S. Sloboda, and *et al.*, "Single seed region growing algorithm in dynamic PET imaging (SSRG/4D-PET) for tumor volume delineation in radiotherapy treatment planning: Theory and simulation," *IEEE Transactions on Nuclear Science*, vol. 59, pp. 2020–2032, 2012. DOI: 10.1109/TNS.2012.2212723.

[127] Y. Tian, L. Xie, and *et al.*, "Beyond masking: Demystifying token-based pre-training for vision transformers," *Pattern Recognition*, vol. 162, p. 111 386, 2025.

[128] S. K. Vashist, P. B. Luppa, and *et al.*, "Emerging technologies for next-generation point-of-care testing," *Trends in Biotechnology*, vol. 33, pp. 692–705, 2015.

[129] A. Vaswani, N. Shazeer, and *et al.*, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[130] Z. Wang, X. Min, and *et al.*, "SMESwin UNet: Merging CNN and Transformer for medical image segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Springer, 2022, pp. 517–526.

[131] E. Xie, W. Wang, and *et al.*, "SegFormer: Simple and efficient design for semantic segmentation with transformers," *Proceedings of the Advances in Neural Information Processing Systems (NeurIPs)*, vol. 34, pp. 12 077–12 090, 2021.

[132] Z. Xing, T. Ye, and *et al.*, "SegMamba: Long-range sequential modeling Mamba for 3D medical image segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Springer, 2024, pp. 578–588.

[133] Y. Xu, Z. Zhang, and *et al.*, "Evo-ViT: Slow-fast token evolution for dynamic Vision Transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 2022, pp. 2964–2972.

[134] F. Yuan, Z. Zhang, and *et al.*, "An effective CNN and Transformer complementary network for medical image segmentation," *Pattern Recognition*, vol. 136, p. 109 228, 2023.

[135] N. Zhang, S. Liu, Z. Hu, and *et al.*, "Accuracy of virtual surgical planning in two-jaw orthognathic surgery: Comparison of planned and actual results," *Oral surgery, Oral medicine, Oral pathology and Oral radiology*, vol. 122, pp. 143–151, 2016.

[136] P. Zhang, C. Tian, and *et al.*, "Intra-head pruning for Vision Transformers via inter-layer dimension relationship modeling," *Neural Networks*, p. 107 656, 2025.

[137] Y. Zhang, H. Liu, and *et al.*, "Transfuse: Fusing Transformers and CNNs for medical image segmentation," in *Proceedings of the Medical Image Computing and Computer Assisted Intervention (MICCAI)*, Springer, 2021, pp. 14–24.

[138] X. Zhao, P. Zhang, and *et al.*, "D2A U-Net: Automatic segmentation of COVID-19 CT slices based on dual attention and hybrid dilated convolution," *Computers in Biology and Medicine*, vol. 135, p. 104 526, 2021.

[139] L. Zhou, H. Liu, and *et al.*, "Token sparsification for faster medical image segmentation," in *Proceedings of the International Conference on Information Processing in Medical Imaging (IPMI)*, Springer, 2023, pp. 743–754.

[140] Z. Zhou, M. M. Rahman Siddiquee, and *et al.*, "UNet++: A nested *U*-Net architecture for medical image segmentation," in *Proceedings of the Medical Imaging Computing and Computer Assisted Intervention (MICCAI)*, Springer, 2018, pp. 3–11.