Visual Segmentation with Explanations for Medical Decision Making: A CNN Architecture with a Transformer-Attention Extension

by

Mahmoud Khalaf

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2026

© Mahmoud Khalaf 2026

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

Research Presented in Part I: An earlier version of this work was published in Khalaf et al. [20]. I was the primary author of that paper, responsible for the writing, and also developed the architecture, methodology, experiments, and all preliminary research. Dr. Cohen provided some suggestions for the writing and also offered meaningful discussion, feedback, and corrections throughout. Dr. Alencar and Dr. Bentahar reviewed the work and offered corrections. Mr. Islam provided feedback on the discussion.

Research Presented in Part II: An earlier version of this part of the work will be published in [21]. As the first author, I developed the dual encoder architecture, the attention-gated fusion mechanism, all experiments, the ablation study, and the explainability methodology, and provided the initial writeup of the paper. Dr. Cohen provided meaningful corrections and feedback throughout and made suggestions for expanding the writing. Mr. Islam provided feedback throughout. It is noted that Dr. Cohen and Mr. Islam co-authored a section on trust that is not included in this thesis, though it inspired aspects of the explainability discussion presented here. Prof. Bentahar's guidance on some of the research in earlier stages led to his inclusion as a coauthor.

All other chapters are original to the thesis and were not based on any published work.

Abstract

Medical image segmentation is essential for assisting medical professionals in locating anomalies in images. However, the lack of explainability in current segmentation frameworks limits clinical trust and adoption. This thesis presents two complementary frameworks for explainable medical image segmentation, demonstrating that accuracy and interpretability can be achieved simultaneously through innovative architectural design. The first framework introduces an attention-based CNN architecture that generates model-specific visual explanations through spatial attention gates integrated directly into the network. The proposed model achieves a Dice score of 0.8621 on the Kvasir-SEG polyp segmentation dataset, outperforming all evaluated explainable models, while producing attention heatmaps that faithfully reflect the model's decision-making process. The second framework advances this by introducing a dual encoder architecture combining a pretrained ResNet-34 CNN encoder with a pretrained Swin Transformer encoder, fused through a learned directional attention-gated mechanism at multiple scales. The dual encoder achieves a Dice score of 0.907 on Kvasir-SEG, outperforming all single encoder baselines, while generating richer multi-scale visual explanations that reflect the complementary contributions of both encoders. Finally, this thesis outlines a pathway in detail towards a fully multimodal explainability system, integrating textual explanations through SigLIP, Retrieval Augmented Generation, and a Large Language Model alongside the visual heatmaps. We explore future directions regarding breast cancer and echocardiogram segmentation, more specifically applications for ejection fraction computation,and insights gained from trials and errors that led to the innovative designs of the thesis.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr. Robin Cohen, for her guidance, support, care, and dedication throughout my entire journey at the University of Waterloo. There were many hurdles along the way, particularly with respect to my health, and Dr. Cohen was always understanding and helped me streamline everything into a manageable schedule despite the challenges of managing a disability. I am deeply grateful for her patience and commitment to my success.

I would also like to thank my co-supervisor, Dr. Jamal Bentahar, for his guidance, support, and for welcoming me into his research lab, which played a major role in the development of Part Two of this thesis. I would like to thank my committee members, Dr. Paulo Alencar and Dr. Olga Veksler, for their valuable feedback and insights. I would also like to thank Dr. Richard Trefler and Edward Chrzanowski for delivering hard copies of this thesis. Furthermore, I would like to sincerely thank Tammy Vucicevich for her assistance with filing many forms on my behalf and helping me step by step with the Concur system.

Moreover, I would like to thank the NSERC CREATE Responsible AI Group for their support with funding and creating an opportunity for my to pursue my graduate studies.

I would like to thank Saidul Islam for his collaboration and continuous support; without him two of the papers would not have been possible. I would also like to thank my lab mates Amanda Spilkin and Hanae El Mekki for their support and for helping me brainstorm and develop ideas for this research. I am grateful to Kathryn Carbone and Olha Wolch for making me feel at home at the University of Waterloo. I would like to thank Nanda Kishore Sreenivas for his technical support during my teaching assistantships and sick days. Finally, I would like to thank Taite La Grange for helping with grading and for helping balance my workload during my extenuating circumstances.

Dedication

To my parents, for their endless sacrifices, for giving beyond their means, and for pushing me further than I ever thought possible. This is yours as much as it is mine.

Table of Contents

Author's Declaration ii

Statement of Contributions iii

Abstract iv

Acknowledgements v

Dedication vi

List of Figures xiii

List of Tables xiv

1 Introduction 1

1.1 Motivation 1

1.2 Contribution 2

1.3 Thesis Outline 3

2 Background and Related Work 5

2.1 Artificial Intelligence and Machine Learning 5

2.2 Data Augmentation and Training for Deep Learning 5

2.3 Convolutional Neural Networks 6

2.3.1 Residual Learning and ResNet . . . 7
2.3.2 Encoders and Decoders . . . 8
2.4 Medical Image Segmentation . . . 8
2.4.1 Colorecal Cancer and Polyp Segmentation . . . 9
2.5 CNN-Based Segmentation Methods . . . 9
2.5.1 U-Net . . . 9
2.5.2 Attention U-Net . . . 11
2.5.3 DUCK-Net . . . 12
2.6 Transformers . . . 12
2.6.1 Self-Attention . . . 12
2.6.2 Vision Transformers . . . 13
2.6.3 Swin Transformer . . . 14
2.7 Transformer-Based Segmentation Methods . . . 14
2.7.1 TransUNet . . . 14
2.7.2 Polyp-PVT . . . 14
2.7.3 TransFuse . . . 15
2.8 Foundation Models for Medical Segmentation . . . 16
2.8.1 MedSAM . . . 16
2.9 Explainability Methods . . . 16
2.9.1 The Need for Explainability in Medical Imaging . . . 16
2.9.2 Gradient-weighted Class Activation Mapping . . . 16
2.9.3 Grad-CAM++ . . . 17
2.9.4 LIME . . . 17
2.9.5 SHAP . . . 18
2.9.6 Attention as Explanation . . . 18
2.10 Vision-Language Models and Textual Explanation . . . 18
2.10.1 Large Language Models . . . 19
2.10.2 Vision-Language Models . . . 19
2.10.3 SigLIP . . . 20
2.10.4 Retrieval Augmented Generation . . . 20

I Model-Specific Explainability using Attention for Medical Imaging **22**

3 Methodology **23**

3.1 Model Overview . . . 23
3.2 Encoder Design . . . 23
3.2.1 Squeeze and Excitation Residual Blocks . . . 23
3.2.2 Atrous and Separable Convolutions . . . 24
3.3 Spatial Attention Gates . . . 25
3.4 Enhanced Duck Block . . . 25
3.4.1 Explainability via Attention Heatmaps . . . 26
3.4.2 Decoder and Segmentation Head . . . 26

4 Experiments **28**

4.1 Dataset . . . 28
4.1.1 Kvasir-SEG . . . 28
4.1.2 Data Preparation and Augmentation . . . 29
4.2 Evaluation Metrics . . . 29
4.2.1 Dice Coefficient . . . 29
4.2.2 Jaccard Index (IoU) . . . 30
4.2.3 Precision and Recall . . . 30
4.3 Implementation Details . . . 30
4.3.1 Experimental Environment . . . 30
4.3.2 Optimiser and Hyperparameters . . . 31
4.3.3 Loss Function . . . 31
4.3.4 Model Configurations . . . 31

5 Results and Discussion **32**

5.1 Quantitative Results . . . 32
5.2 Qualitative Results and Attention Heatmap Visualisation . . . 33

6 Discussion **35**

6.1 Interpretation of Results . . . 35

6.2 Comparison with Post-hoc Explainability Methods . . . 36

6.3 Limitations . . . 36

II Dual Encoder Explainable Medical Image Segmentation via Attention-Gated Fusion **37**

7 Methodology **38**

7.1 Model Overview . . . 38

7.2 Spatial Feature Encoding . . . 39

7.2.1 ResNet34 Encoder . . . 39

7.3 Global Feature Encoding . . . 40

7.3.1 Swin Transformer-Base Encoder . . . 40

7.3.2 Channel Projection . . . 40

7.4 Dual Encoder Attention Fusion . . . 41

7.4.1 Attention Gate Mechanism . . . 41

7.4.2 Swin-Guided ResNet Attention . . . 41

7.4.3 Multi-Scale Fusion . . . 42

7.5 Decoder and Segmentation Head . . . 42

7.5.1 Decoder Blocks . . . 42

7.5.2 Final Upsampling . . . 42

7.5.3 Segmentation Output . . . 42

7.6 Explainability via Attention Weights . . . 43

7.6.1 Attention Weight Extraction . . . 43

7.6.2 Visualization . . . 43

8 Experiments **44**
8.1 Data . . . 44
8.1.1 Kvasir-SEG . . . 44
8.1.2 Data Augmentation . . . 44
8.2 Metrics . . . 45
8.2.1 Dice Coefficient . . . 45
8.2.2 Jaccard Index . . . 45
8.2.3 Binary Cross Entropy Loss . . . 45
8.3 Implementation Details . . . 46
8.3.1 Experiment Environment . . . 46
8.3.2 Loss Function . . . 46
8.3.3 Optimiser and Learning Rate . . . 46
8.3.4 Model Configuration . . . 46

9 Results and Discussion **48**
9.1 Model Performance . . . 48
9.1.1 Quantitative Results . . . 48
9.1.2 Qualitative Results . . . 49
9.2 Ablation Study . . . 49
9.2.1 Skip Connections . . . 49
9.2.2 Attention Guidance and Parameter Freezing . . . 50
9.3 Comparison with State of the Art . . . 51
9.4 Explainability Analysis . . . 52
9.4.1 Attention Weight Visualisation . . . 52
9.4.2 Comparison with Post-hoc Methods . . . 52
9.5 Discussion . . . 53
9.5.1 Interpretation of Results . . . 53
9.5.2 Limitations . . . 53

III Contributions of Thesis and Future Work Towards Multimodal Explainable AI **54**

10 Reflections and Contributions **55**

10.1 Summary of Contributions 55

10.2 Reflection on Model-Specific Explainability 56

10.3 Relationship to Prior Work 57

10.4 Looking Ahead 57

11 Future Work and Conclusion **59**

11.1 Multimodal Textual Explanation via RAG and LLM 59

11.2 Echocardiogram Segmentation and Ejection Fraction Computation 61

11.2.1 Dataset 61

11.2.2 The Preliminary ViT Architecture 62

11.2.3 Challenges and Findings 62

11.2.4 LLM Prototype and Textual Explanation 63

11.2.5 Future Directions 66

11.2.6 Breast Cancer Segmentation 67

11.2.7 Attention Map Aggregation 68

11.2.8 Recommendations for Next Generation of Medical Datasets 68

11.2.9 Being Attuned to Medical Experts 69

11.3 Conclusion 69

References **71**

List of Figures

2.1 CNN Architecture [2] 7
2.2 U-Net Architecture [35] 10
2.3 Attention Gate [31] 12
2.4 Vision Transformer [12] 13
2.5 TransUNet [7] 15
2.6 SigLIP [49] 20

3.1 Enhanced Explainable Duck Block 26

4.1 Kvasir-SEG Dataset Sample 28

5.1 Multi-scale Attention Heatmaps 33

7.1 Dual Encoder 39

9.1 Dual Encoder Attention Heatmaps 52

11.1 Dual Encoder + LLM 60
11.2 CAMUS Sample 61
11.3 Example of patch-level segmentation artifacts produced by the ViT encoder. Each white square represents a 16×16 patch predicted as left ventricle, resulting in blocky boundaries unsuitable for precise cardiac structure delineation. 63
11.4 Example of Left Ventricle Segmentation 66

List of Tables

5.1 Architecture comparisons and model support of explanation on Kvasir-SEG 32

9.1 Performance comparison across different training strategies 48

9.2 Performance comparison of pretrained models with and without skip connections 50

9.3 Performance comparison of Attention Guidance & Parameter Freezing in dual encoder architectures 50

9.4 Performance comparison of best models across different architectures . . . 51

Chapter 1

Introduction

1.1 Motivation

Computer vision methods for segmenting visual images can be of considerable assistance for medical professionals to interpret scans and other images, towards their decision making for patients. In this thesis, we explore how to carry out this processing in a way that best supports providing explanations of the accompanying AI system to these medical experts. The motivation is to provide an effective human-AI partnership that may lead to better outcomes for this important real-world application. We are also invested in advancing research in explainable AI and in the use of visual segmentation for medical applications.

The clinical motivation for this work is grounded in the problem of colorectal cancer (CRC) detection. CRC is the third most common cancer worldwide and the second leading cause of cancer-related deaths. CRC primarily develops from adenomatous polyps, which can be effectively treated if identified and removed at an early stage [40]. Automated polyp segmentation from colonoscopy images therefore represents a high-impact application of AI in healthcare, offering the potential to enhance clinical workflows and improve patient outcomes. However, the deployment of AI systems in clinical settings requires not only accurate segmentation, but also clear and interpretable explanations of why target regions are delineated, in order to establish the trust of medical professionals and to support responsible clinical decision making.

1.2 Contribution

This thesis makes contributions across two core models for explainable medical image segmentation, with a third part that reflects on extensions and future directions.

With respect to computer vision methods for segmenting images, our work is grounded in the use of convolutional neural networks (CNNs) and our specific focus, to begin with, is studying distinct architectures for carrying out this processing. We conduct an extensive study of architectures that are suitably efficient and accurate, and then move on to discuss an approach to enable these solutions to properly support explanations as well. The framework that we develop is described in detail in Part One of this thesis, together with results which show that it is possible to be both accurate and explainable. A key insight of Part One is that model-specific explanations, derived directly from the attention mechanisms integrated into the architecture, are more faithful to the model's actual decision-making process than model-agnostic post-hoc methods such as GradCAM [38], which are often adopted by researchers for medical image interpretation tasks.

We then move forward from this first model to examine how to improve upon the visual processing using a design based on transformers and attention, and to do so in such a way that explanations are not only supported but also enhanced for the medical expert end users. This second model, presented in Part Two, employs a dual encoder architecture that combines a pretrained ResNet-34 CNN encoder with a pretrained Swin Transformer encoder, fusing their complementary representations through a learned attention-gated mechanism at multiple scales. The CNN encoder captures fine-grained local spatial features, while the Swin Transformer encoder captures global contextual information across the entire image. The attention-gated fusion mechanism selectively combines these representations based on their relevance to the segmentation task, and as a byproduct generates multi-scale visual explanations that provide clinicians with a richer and more informative view of the model's decision-making process than was possible in Part One. Results of experiments run with this new model demonstrate its effectiveness in terms of accuracy, outperforming all single encoder baselines when evaluated on the Kvasir-SEG[18] medical image dataset.

Motivated by extending the explanations offered to users to also integrate textual explanations alongside visual ones that highlight regions of interest, Part Three of this thesis discusses important extensions for future work. One significant step forward would make use of Large Language Models and Retrieval Augmented Generation in order to yield a richer depiction of the processing of the AI system, and more extensive human-AI partnerships. We clarify how the dual encoder model designed in Part Two supports the extension

into this richer hybrid output, as a further argument for why our research has great potential. We also discuss additional work completed that examined how to work in partnership with medical experts in a domain distinct from the one used to illustrate our methods in Part One and Part Two, namely that of cardiology, providing deeper insights into why the methods we integrate have promise for extended settings.

We also pause in Part Three to reflect on the novelty of our work on explainable AI compared to that of other researchers, and to clarify our relationship to other work on visual segmentation. For this latter topic, we reflect on three main steps taken: (i) the adoption of a U-Net architecture with attention mechanisms to support model-specific explainability, (ii) the integration of a Swin Transformer encoder alongside a CNN encoder in a dual encoder design, and (iii) an improvement upon prior dual-stream designs such as TransFuse, in a way that better supports the important consideration of explainable AI through directional attention-gated fusion and model-specific multi-scale heatmaps.

1.3 Thesis Outline

Chapter 2 presents Background and Related Work, introducing technical concepts and research that inform the models and results presented throughout the thesis, including CNN architectures, vision transformers, attention mechanisms, and explainability methods.

Part One introduces the first model developed in this thesis: an attention-based CNN architecture for explainable medical image segmentation. Chapter 3 presents the methodology, detailing the proposed architecture and its attention-based explainability mechanism. Chapter 4 presents the experimental setup, including the dataset, evaluation metrics, and implementation details. Chapter 5 presents the results and discussion, demonstrating the value of the proposed design for accurate and explainable polyp segmentation. Chapter 6 provides an interpretation of the results and briefly points to limitations.

Part Two introduces the second model: a dual encoder architecture combining a CNN encoder and a Swin Transformer encoder with attention-gated fusion. Chapter 7 presents the methodology, detailing the dual encoder design and the attention-gated fusion mechanism. Chapter 8 presents the experimental setup. Chapter 9 presents the results and discussion, demonstrating the effectiveness of the dual encoder architecture in terms of both segmentation accuracy and rich visual explanations produced. This chapter also draws out the potential usefulness of attention maps as an element that may assist clinicians in understanding and trusting results.

Part Three presents the closing contributions of this thesis. The chapter provides an

overall reflection on the contributions made across both parts, a discussion of the novelty of the explainability approach relative to prior work, and an elaboration of important directions for future work including textual explanation integration and extensions to additional medical domains. Chapter 10 provides a concise conclusion to the thesis. Chapter 11 then discusses valued avenues for future work, describing first some steps already taken with applications to medical concerns beyond polyp detection.

We are encouraged that this work encompasses the central concern of offering new methods and models for designing AI systems, but in ways that can be trusted and understood by the medical professionals that rely on them. In so doing, we are able to highlight the social value of AI research.

Chapter 2

Background and Related Work

2.1 Artificial Intelligence and Machine Learning

Artificial Intelligence (AI) is a specific field involving the development of computer systems capable of performing tasks that typically require human intelligence [37]. These tasks range from recognizing patterns, helping users to make decisions, to understanding natural language. Machine learning is a field within AI that enables a system to learn from data instead of being explicitly programmed [37]. Models trained on data improve their performance through each training iteration. Computer vision and natural language processing have witnessed a large expansion due to abundant curation of datasets for various applications. Deep learning uses neural networks with many layers to learn representations of data. The models can vary in size and depth, allowing many abstract features to be learned through layers in the architecture. Within the field of computer vision, tasks such as image classification, object detection and image segmentation have been enriched with the use of deep learning. The work presented in this thesis falls within the domain of AI, specifically focusing on deep learning methods for medical image segmentation, as well as the concern of AI explainability.

2.2 Data Augmentation and Training for Deep Learning

As deep learning models grow large and more complex, they are able to capture data representations immaculately. However, a key challenge in training deep learning models

(adjusting model parameters) is overfitting. Overfitting is when the model memorizes the training data rather than generalizing to unseen test data. This can be an issue for medical images, where datasets may be small in size, with the difficulty in curating datasets due to patient privacy laws and consent. Data augmentation addresses this by artificially expanding the training dataset by applying random transformations on the images such as rotations, flips, brightness, contrast adjustments, zoom, and blurring. These transformations introduce variability into the training data, prompting the model to learn features rather than memorizing specific images and their desired output.

As a model learns to map an input to the desired output, it is presented with a large dataset of samples, and uses each sample to make predictions, computing a loss to measure how far the predictions are from the ground truth. A backpropagation process, repeated over many epochs, can help to minimize the loss, until the model converges to a solution that generalizes to unseen data. An optimizer is the algorithm used when performing backpropagation in order to adapt the learning rate progressively. The most widely used optimizer in deep learning is Adam [22], and this is the one that is used within the implementations of the thesis.

2.3 Convolutional Neural Networks

CNNs have revolutionized image classification tasks, achieving cutting-edge results in object detection. The success of CNNs is largely attributed to their ability to hierarchically extract and learn features from unstructured image data [10].

The CNN architecture displayed in Figure 2.1 consists of four major components:

- **Convolutional Layers**: These layers apply convolution filters in the form of kernels onto an input image, sliding over the image to extract patterns such as edges, corners, and textures. Early convolutional layers learn to detect low-level features such as edges, while deeper layers learn higher-level features such as shapes or facial expressions.

- **Pooling Layers**: Convolutional layers are often followed by pooling layers that reduce the spatial dimensions of the learned feature maps, improving computational efficiency and reducing overfitting. Max pooling selects the maximum value within a defined window; for example, a 64×64 feature map is reduced to 32×32 when passed through a 2×2 max pooling layer.

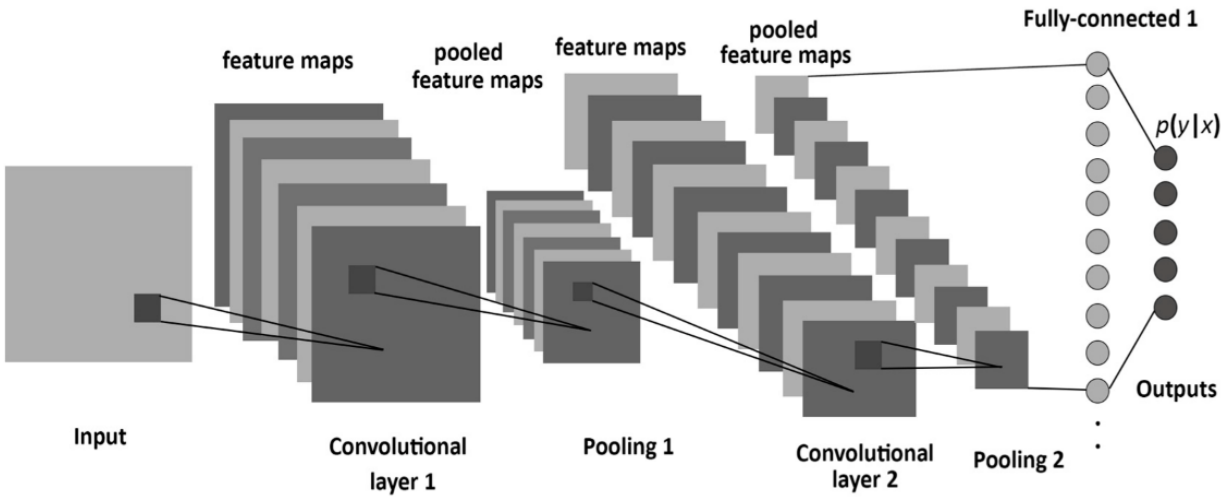

Figure 2.1: CNN Architecture [2]

- **Activation Functions**: Non-linear activation functions are used throughout neural networks to allow the model to learn complex relationships between inputs and outputs. The most commonly used activation function in CNNs is the Rectified Linear Unit (ReLU), typically applied after every convolutional layer.
- **Fully Connected Layers**: At the end of the network, fully connected layers combine all extracted features to perform classification, mapping learned representations to output neurons. Sigmoid activation is used for binary classification, while softmax is used for multi-class tasks.

Models such as AlexNet [23], VGG [41], and ResNet [16] have attained remarkable results on the ImageNet [36] benchmark. However, classification as a utility is limited for medical imaging, as classification models predict categories such as presence or absence of a condition, which is insufficient for clinical applications that require precise localization.

2.3.1 Residual Learning and ResNet

During training, gradients are the values that tell a neural network how much to adjust its weights. A key challenge in training deep CNN architectures is the vanishing gradient problem, where gradients become increasingly small as they propagate back through many layers, causing early layers to stop learning effectively. This limits the depth of networks

that can be effectively trained. He et al. [16] address this through residual learning, introducing skip connections within each block that allow gradients to flow directly through the network. Each residual block computes the following:

$$y = \mathcal{F}(x, W) + x \tag{2.1}$$

where x is the input to the block, W represents the learnable weights, and $\mathcal{F}(x, W)$ is the residual mapping to be learned. The resulting ResNet architecture has become a foundational backbone for many downstream computer vision tasks, and serves as the CNN encoder in the dual encoder architecture proposed in Part Two of this thesis.

2.3.2 Encoders and Decoders

An encoder-decoder architecture is the most prevalent method for image segmentation tasks. An encoder process an input image and extracts meaningful features by down sampling the images through multiple convolution or attention layers. This results in a compression of the dimensions of the image while increasing feature channels. Low-level features such as edges are captured in the early stages of the encoding, while high-level semantic information in the deeper layers. The decoder decompresses the encoding representations and upsamples to the original image resolution, producing a pixel-wise segmentation mask or reconstruction of the original image.

2.4 Medical Image Segmentation

Deep learning neural networks have achieved state-of-the-art performance for many visual recognition tasks [26]. In particular, convolutional neural networks classify an input image into a single output class label. However, classifying medical images into a single class may not be sufficiently useful for health professionals. An ideal output should assign a class label to each specific pixel of an input image, resulting in a pixel-wise classification known as image segmentation.

Medical image segmentation is the process of partitioning a medical image into regions, enabling the precise locations and delineation of anatomical structures or anomalies. For medical professionals to determine whether there is concern with a particular patient, anomalies in images become particularly important to detect, with the location, shape or size of the identified anomalies also becoming relevant. Accurate image segmentation is

essential in clinical deployment for a wide range of diagnostic and treatment planning, specifically tumor detection, organ delineation, and surgical planning [26]. Deep learning tools that automatically detect and flag anomalies in clinical workflows may enhance diagnostic accuracy and improve healthcare accessibility and patient outcomes.

2.4.1 Colorecal Cancer and Polyp Segmentation

Colorectal cancer (CRC) is the third most common cancer worldwide and the second leading cause of cancer mortality [40]. CRC primarily develops as adenomatous polyps, which have the potential to become cancerous, especially if they are large and show cellular changes such as dysplasia. Adenomatous polyps therefore serve as biological markers for potenital CRC, making early detection and removal imperative for effective prevention and treatment [19]. The challenge of polyp segmentation arises from high variability in polyp appearance, size, shape, and location. Additionally, visual similarity between polyps and background tissue, make it difficult for segmentation tasks to delineate the difference. We have chosen to showcase our proposed designs in this thesis for this particular medical application.

2.5 CNN-Based Segmentation Methods

CNNs have been applied to analyze medical images, achieving physician-level accuracy [26]. In this section we outline some of the models that have gained considerable importance and which motivate the design of our architecture.

2.5.1 U-Net

U-Net [35] (Figure 2.2) introduces the foundational encoder-decoder architecture for medical image segmentation. A contracting CNN path progressively downsamples the input image into compact feature representations, while a symmetric expansive path upsamples these representations back to the original resolution.

The network consists of a compression path (encoder) and expansive path (decoder) connected through skip connections at each parallel stage, forming a U- shape. The encoding path, consists of double 3×3 convolutions followed by a batch normalization and a rectified linear unit (ReLU) activation. A 2×2 max-pooling layer with strides of 2 down

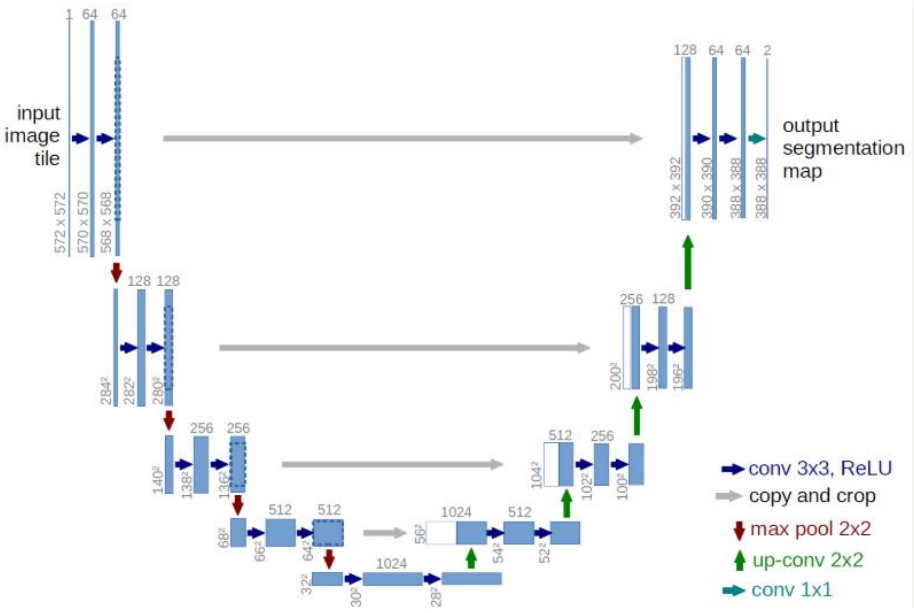

Figure 2.2: U-Net Architecture [35]

samples each double convolution into the next stage. At each down-sampling step, the number of features is doubled, starting from 64 filters to 512 at the bridge/bottleneck.

The decoding path consists of upsampling operations that half the the features and upscale the image, essentially undoing the encoding operations. Each upsampling operation is similarly followed by a double 3×3 convolution, until the original feature map size is reached. The skip connection is a concatenation of the parallel encoding feature map that merges low-level and high-level features at each stage. The final layer utilizes a 1×1 convolution to map the final feature map to the number of desired classes.In the case of polyp segmentation, the output is binary (i.e. polyp exists or does not exist), thus a sigmoid activation function is used at the output, predicting values of 0 or 1.

The basic operations in U-Net can be expressed mathematically. For the contracting path, each convolution block performs:

$$f_l = \text{ReLU}(\text{BN}(\text{Conv3} \times 3(xl - 1))) \tag{2.2}$$

The max pooling operation for downsampling:

$$x_l = \text{MaxPool}_{2\times2}(f_l) \tag{2.3}$$

In the expansive path, the upsampling and concatenation operations are:

$$u_l = \text{UpConv2} \times 2(xl + 1) \tag{2.4}$$

$$x_l = \text{Concat}[f_l, u_l] \tag{2.5}$$

The final prediction for pixel classification:

$$y = \sigma(\text{Conv}_{1\times1}(x_L)) \tag{2.6}$$

where σ represents the sigmoid activation function.

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{2.7}$$

The key contribution of U-Net is the introduction of skip connections, which concatenate encoder feature maps directly to their corresponding decoder layers, preserving spatial detail that is otherwise lost during downsampling.

2.5.2 Attention U-Net

Attention U-Net [31] extends the original U-Net by replacing plain skip connections with attention gates. Before a skip connection is passed to the decoder, an attention gate computes a spatial mask using a gating signal from the deeper decoder layer and the skip feature map. This mask suppresses irrelevant background regions, allowing the model to focus on the target anatomy without relying on supervision. The attention mechanism generates and propagates relevant features at each stage through the skip connections.

Each attention gate, displayed in Figure 2.3, takes two inputs: A feature map from the encoder (skip connection) and a gating signal from the decoder. The gating signal is used to produce attention coefficients. The coefficients are directly applied to the skip connections before concatenation with the decoder features.

The attention gate generates attention coefficients α through:

$$\alpha = \sigma\left(\text{Conv}_{1\times1}\left(\text{ReLU}\left(\text{BN}\left(\text{Conv}_{1\times1}(g)\right) + \text{BN}\left(\text{Conv}_{1\times1}(x)\right)\right)\right)\right) \tag{2.8}$$

where x represents features from skip connection and g is the gating signal. The refined features after attention is:

$$\hat{x} = x \cdot \alpha \tag{2.9}$$

The final output after concatenation is:

$$y_l = \text{Concat}[\hat{x}, g] \tag{2.10}$$

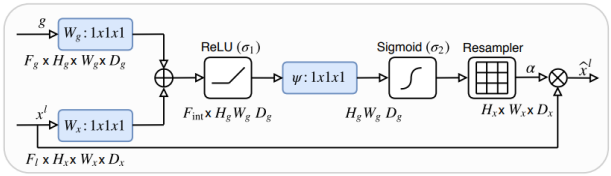

Figure 2.3: Attention Gate [31]

2.5.3 DUCK-Net

DUCK-Net [13] is a CNN-only architecture designed specifically for polyp segmentation that introduces the DUCK block, a multi-scale convolutional module that processes features at different receptive field sizes in parallel before merging them into a single representation. The DUCK block operation consists of 6 branches of various convolutional blocks in parallel. The network optimizes the weights when training and selects the most optimal branches automatically. This design allows the model to capture both fine-grained local texture and structural patterns within a single convolutional stage.

DUCK-Net achieves competitive performance on Kvasir-SEG[18] and CVC-ClinicDB[3] benchmarks without any transformer component. It serves as a strong CNN baseline against which the proposed dual-encoder design is compared, with the hypothesis that incorporating global context through a Swin Transformer encoder addresses the inherent receptive field limitations of convolutional approaches. However, the DUCK-Net architecture does not support explainability, providing no mechanism for generating visual explanations of its segmentation decisions, directly motivating the work presented in Part One of this thesis.

2.6 Transformers

2.6.1 Self-Attention

Vaswani et al. [45] introduce the Transformer architecture, originally developed for natural language processing tasks. The self-attention mechanism computes relationships between all elements of an input sequence simultaneously, enabling the capture of long-range dependencies without the locality constraints imposed by convolutional kernels. For an input sequence, the self-attention mechanism computes queries Q, keys K, and values V, producing an output as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{2.11}$$

where d_k is the dimension of the keys, used to scale the dot products and prevent vanishing gradients in the softmax function.

2.6.2 Vision Transformers

Dosovitskiy et al. [12] adapt the Transformer architecture to computer vision in the form of the Vision Transformer (ViT). The ViT in Figure 2.4 divides an input image into fixed-size patches, typically 16×16 pixels, which are projected into embeddings compatible with the Transformer input. This approach allows the architecture to capture long-range spatial relationships and global context across the entire image using self-attention mechanisms. However, ViT presents significant computational challenges, as training from scratch requires large datasets and substantial computational resources due to the quadratic complexity of self-attention operations. Furthermore, the fixed-size patch division may result in coarse segmentations along object boundaries, limiting precise boundary localization.

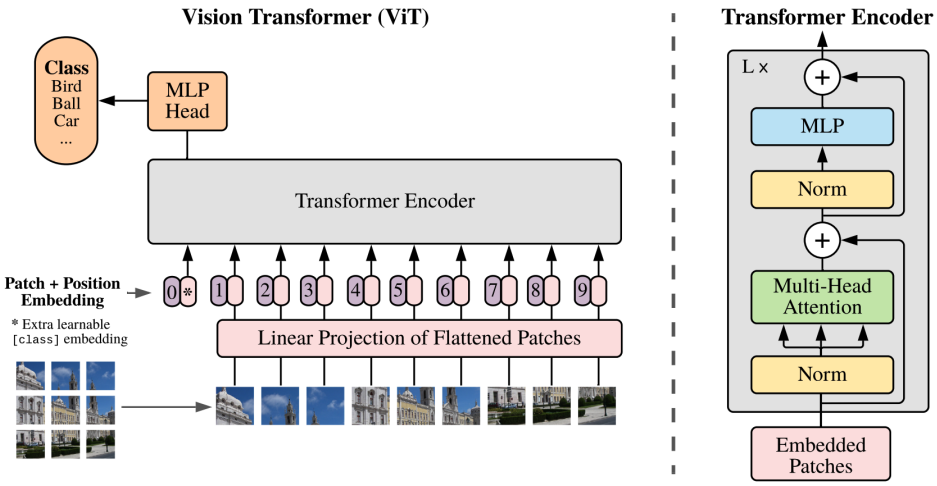

Figure 2.4: Vision Transformer [12]

2.6.3 Swin Transformer

Liu et al. [27] introduce the Swin Transformer, which addresses the computational limitations of ViT by computing self-attention within local windows rather than globally across the entire image. The shifted window mechanism allows the model to capture both local and global contextual information while significantly reducing the computational complexity of self-attention from quadratic to linear with respect to image size. The Swin Transformer also produces hierarchical feature representations at multiple scales, making it well suited for dense prediction tasks such as segmentation. These properties make the Swin Transformer a more practical choice than the original ViT for medical image segmentation tasks, and it serves as the transformer encoder in the dual encoder architecture proposed in Part Two of this thesis.

2.7 Transformer-Based Segmentation Methods

2.7.1 TransUNet

TransUNet [7] was among the first architectures (see Figure 2.5) to integrate a Vision Transformer into a medical image segmentation pipeline. A ResNet encoder extracts convolutional feature maps which are then flattened into patch tokens and processed by a ViT to model long-range spatial dependencies. The resulting tokens are decoded using a UNet-style upsampling path with skip connections. A limitation relevant to this work is that the transformer operates only at the bottleneck level, meaning the shallower encoder stages remain purely convolutional with no global context modelling. In contrast, the proposed architecture in Part Two employs a Swin Transformer encoder that produces hierarchical multi-scale features at all four encoding stages, providing global context throughout the entire encoding pathway.

2.7.2 Polyp-PVT

Polyp-PVT [5] adopts the Pyramid Vision Transformer (PVT) as a pure transformer backbone, replacing the CNN encoder entirely. PVT produces hierarchical multi-scale feature maps natively, which are fed into a cascaded decoder with progressive feature fusion. The model achieves strong results on standard polyp segmentation benchmarks including Kvasir-SEG. However, by discarding convolutional encoders entirely, Polyp-PVT forgoes the fine-grained local texture features that CNNs capture through inductive spatial biases.

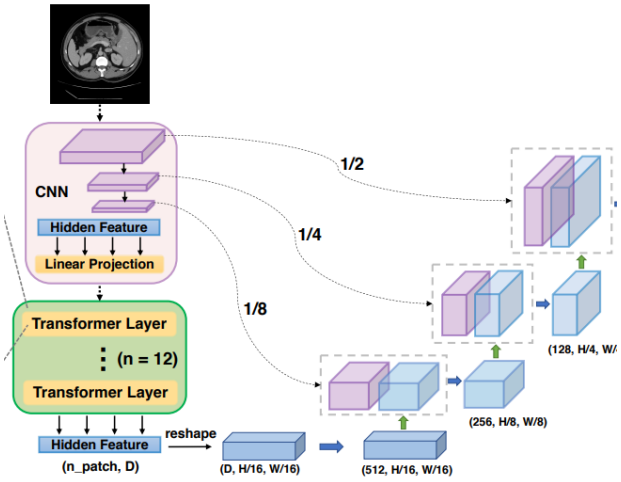

Figure 2.5: TransUNet [7]

The proposed architecture addresses this by retaining a ResNet-34 encoder in parallel with the Swin Transformer, combining local CNN texture features with global transformer context through an attention-based fusion mechanism.

2.7.3 TransFuse

TransFuse [51] is the most closely related prior work to the architecture proposed in this thesis. It processes the input image through a ResNet and a Vision Transformer in parallel, then fuses their feature maps at multiple scales using a BiFusion module that combines additive attention with element-wise multiplication. This dual-stream design demonstrates that CNN and transformer representations are complementary and that their fusion consistently outperforms either encoder alone on polyp segmentation tasks. The proposed architecture in Part Two follows this same parallel dual-encoder paradigm but differs in two key respects. First, the transformer component is a Swin Transformer-Base rather than a standard ViT, providing hierarchical multi-scale features with linear complexity. Second, the fusion strategy employs a directional attention gate in which the Swin features explicitly guide what spatial regions to attend to within the ResNet feature maps, rather than symmetric additive fusion.

2.8 Foundation Models for Medical Segmentation

2.8.1 MedSAM

MedSAM [29] adapts the Segment Anything Model (SAM) to the medical imaging domain by fine-tuning it on a large and diverse corpus of medical image-mask pairs spanning multiple modalities and anatomical regions. SAM employs a Vision Transformer encoder (ViT-H) paired with a prompt-based decoder that accepts spatial inputs such as bounding boxes or point coordinates to indicate the target region of interest. While MedSAM demonstrates strong generalisation across medical imaging tasks, it is fundamentally a prompted segmentation system and is therefore not fully automated. At inference time, a human-provided spatial prompt is required to indicate the region to segment. The architecture proposed in this thesis operates without any such prompt, taking only the raw endoscopic image as input and producing a segmentation mask end-to-end, making it more suitable for real-time clinical deployment in automated colonoscopy analysis pipelines.

2.9 Explainability Methods

2.9.1 The Need for Explainability in Medical Imaging

The black-box nature of deep learning models in medical image segmentation poses a significant challenge in clinical settings where interpretability and trust are imperative. Medical professionals require not only accurate segmentations but also clear explanations of why target regions are delineated in order to validate model outputs and make informed clinical decisions. The lack of explainability in current medical image segmentation frameworks demonstrates a gap in assisting clinicians in understanding how segmentation decisions are made, motivating the model-specific explainability approaches proposed in this thesis.

2.9.2 Gradient-weighted Class Activation Mapping

Gradient-weighted Class Activation Mapping (Grad-CAM) [38] produces a coarse localisation map highlighting the regions of an input image most influential to a model's prediction. It does so by computing the gradient of the target class score with respect to the final convolutional feature maps, then performing a weighted average of those feature maps using the gradients as importance weights. The resulting heatmap is upsampled to the input

resolution and overlaid on the original image. Grad-CAM requires no architectural modification and can be applied to any CNN post-hoc, making it widely adopted for visualising medical image classifiers and segmentation models. However, because it operates on a single convolutional layer and relies on global average pooling of gradients, it tends to produce spatially coarse maps that may miss fine-grained boundaries, which is a known limitation in polyp segmentation where precise boundary localisation is clinically important.

2.9.3 Grad-CAM++

Grad-CAM++ [6] extends Grad-CAM by replacing the global average of gradients with a pixel-wise weighting scheme that accounts for the partial second-order derivatives of the class score with respect to the feature maps. This produces importance weights that more accurately reflect the contribution of individual spatial locations within a feature map, rather than treating each activation channel uniformly. In practice, Grad-CAM++ generates sharper and more spatially precise saliency maps than Grad-CAM, particularly when multiple instances of the target object are present or when the object occupies a small region of the image. Despite this improvement, Grad-CAM++ remains a post-hoc gradient-based method and inherits the same fundamental limitation: the explanation is derived after the prediction is made and is not guaranteed to faithfully reflect the internal decision process of the model.

2.9.4 LIME

Local Interpretable Model-agnostic Explanations (LIME) [34] explains individual predictions by approximating the behaviour of a complex model locally with a simpler, interpretable surrogate. For image inputs, LIME segments the image into superpixels, generates a set of perturbed samples by randomly occluding subsets of superpixels, and observes how the model's output changes across these perturbations. A weighted linear model is then fitted to these samples to identify which superpixels contributed most to the prediction. LIME is model-agnostic and can be applied to any black-box model without access to gradients or internal representations. Its primary limitation is computational cost, as each explanation requires hundreds of forward passes through the model. Additionally, the quality of the explanation is sensitive to the superpixel segmentation and the number of perturbation samples, introducing instability across repeated runs on the same input.

2.9.5 SHAP

SHapley Additive exPlanations (SHAP) [28] provides a unified framework for feature attribution grounded in cooperative game theory. Each feature is assigned a Shapley value representing its average marginal contribution to the model output across all possible feature subsets. For deep learning models, DeepSHAP combines SHAP values with backpropagation to efficiently approximate these attributions, producing pixel-level importance maps that satisfy desirable theoretical properties including local accuracy, consistency, and missingness. SHAP values are considered among the most theoretically principled post-hoc attribution methods available. However, exact Shapley value computation is exponentially expensive, and practical approximations introduce error. Like Grad-CAM and LIME, SHAP operates entirely post-hoc and produces explanations that are decoupled from the model's internal reasoning process.

2.9.6 Attention as Explanation

In contrast to post-hoc methods, model-specific explainability approaches generate explanations as an inherent part of the model's forward pass, ensuring that the explanations are faithful to the model's actual decision-making process. Attention mechanisms provide a natural avenue for model-specific explainability in segmentation models, as the learned attention weights directly reflect which spatial locations and features the model considers most relevant at each stage of the prediction. The spatial attention coefficients computed during the forward pass can be visualized as heatmaps and overlaid on the original image, providing clinicians with interpretable visual cues that are grounded in the model's actual reasoning rather than a post-hoc approximation. This approach contrasts with model-agnostic methods such as GradCAM and LIME, which may produce explanations that are inconsistent with the model's internal decision process, and is the foundation of the explainability approach proposed in both Part One and Part Two of this thesis.

2.10 Vision-Language Models and Textual Explanation

The integration of vision and language has emerged as a promising direction for medical AI, enabling systems to not only process visual inputs but also generate natural language descriptions of their findings. This section introduces the key components that underpin the textual explanation framework proposed as future work in Part Three of this thesis.

2.10.1 Large Language Models

Large Language Models (LLMs) are neural networks trained on massive text corpora that have demonstrated remarkable ability to understand and generate human language [45]. Models such as GPT [1] and Gemma [43] are capable of generating coherent, contextually relevant text from a given prompt. In medical contexts, LLMs have shown promise for tasks such as clinical report generation, question answering, and summarizing medical findings [30]. However, LLMs rely on parametric knowledge encoded during training, which may be outdated or insufficiently specific for niche clinical contexts. This limitation motivates the use of Retrieval Augmented Generation, discussed below, to ground LLM outputs in up-to-date domain-specific medical knowledge.

Chain of Thought

Chain of Thought prompting [47] is a method for improving model reasoning of LLMs. The chain of thought forces the model to produce reasoning steps before arriving at its answer, which is essentially a representation of how the model thinks. Instead of prompting for an output, chain of thought prompts for instructions and thinks through the problem step by step, making the reasoning process transparent. This method is particularly valuable in medical decision making, where the reasoning behind a conclusion is as important as the conclusion itself.

2.10.2 Vision-Language Models

Vision-Language Models (VLMs) extend the capabilities of LLMs by jointly processing visual and textual inputs, enabling models to reason about images and generate natural language descriptions of their content [50]. VLMs are typically trained on large datasets of image-text pairs, learning to align visual and linguistic representations in a shared embedding space. Notable examples include CLIP [33], which aligns image and text embeddings through contrastive learning, and PaliGemma [4], which combines a powerful vision encoder with a language model to enable multimodal understanding and generation. In medical imaging, VLMs have shown promise for tasks such as radiology report generation, visual question answering, and medical image captioning [15]. However, most VLMs do not natively support dense prediction tasks such as segmentation, limiting their direct applicability for pixel-wise medical image analysis. The framework proposed in Part Three addresses this by integrating a VLM component alongside the dual encoder segmentation architecture, rather than replacing it.

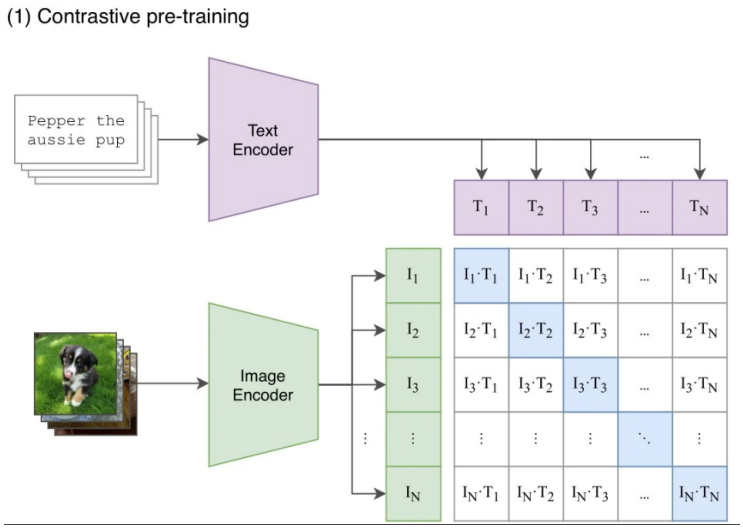

Figure 2.6: SigLIP [49]

2.10.3 SigLIP

SigLIP (Sigmoid Loss for Language-Image Pre-Training, see Figure 2.6) [49] is a vision-language model that improves upon the contrastive learning approach of CLIP [33] by replacing the softmax normalization with a pairwise sigmoid loss. This modification removes the requirement for a global view of pairwise similarities across the batch, enabling better scalability and improved performance on image-text retrieval tasks. SigLIP produces strong image-text alignment, making it well suited for generating semantically meaningful text queries from image embeddings. In the proposed future work, the Swin Transformer embeddings from the dual encoder architecture are passed to SigLIP to generate text queries that describe the identified anomaly, bridging the gap between visual segmentation and natural language explanation.

2.10.4 Retrieval Augmented Generation

Retrieval Augmented Generation (RAG) [25] addresses the limitation of LLMs relying solely on parametric knowledge by augmenting the generation process with relevant information retrieved from an external knowledge base. Given a query, the RAG system retrieves the most relevant documents from the knowledge base using a similarity search, and provides them as additional context to the LLM alongside the original query. This

grounds the generated output in specific, up-to-date knowledge rather than relying on what the model learned during training. In medical applications, RAG is particularly valuable as clinical knowledge evolves rapidly and accuracy of information is critical [48]. In the proposed future work, the queries generated by SigLIP are passed to a RAG system that retrieves relevant clinical literature and medical guidelines specific to the identified anomaly, ensuring that the textual explanations generated by the LLM are grounded in domain-specific medical knowledge.

Part I

Model-Specific Explainability using Attention for Medical Imaging

*An earlier version of this work appeared in [20]

Chapter 3

Methodology

3.1 Model Overview

This chapter presents the proposed architecture for explainable medical image segmentation. The model performs accurate polyp segmentation while simultaneously generating model-specific visual explanations in the form of attention heatmaps. The architecture builds upon the U-Net encoder–decoder backbone, integrating key components such as Squeeze-and-Excitation (SE) residual blocks, spatial attention gates, and an enhanced parallel convolution block adapted from DUCK-Net. The result is a model that achieves competitive segmentation performance while supporting interpretable, model-specific explanations aligned with the segmentation process.

3.2 Encoder Design

3.2.1 Squeeze and Excitation Residual Blocks

The proposed model expands upon the base DUCK-Net residual blocks by incorporating Squeeze and Excitation (SE) blocks [17]. SE blocks adaptively recalibrate channel-wise feature responses by modeling interdependencies between channels. The squeeze operation aggregates feature maps across spatial dimensions ($H \times W$) using global average pooling to produce a channel descriptor:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j) \tag{3.1}$$

where $u_c(i, j)$ represents the input feature map at position (i, j) for channel c. The excitation operation then captures channel-wise dependencies through a gating mechanism with sigmoid activation, implemented using two fully connected layers with a bottleneck design:

$$s = \sigma(W_2 \text{ReLU}(W_1 z)) \tag{3.2}$$

$$W_1 \in \mathbb{R}^{\frac{C}{r} \times C}, \qquad W_2 \in \mathbb{R}^{C \times \frac{C}{r}} \tag{3.3}$$

where C denotes the number of feature channels, σ is the sigmoid activation function, and W_1, W_2 are learned parameters. The reduction ratio r controls the bottleneck compression in the excitation module by reducing the channel dimensionality from C to $\frac{C}{r}$ before restoring it back to C, enabling the network to reduce computational complexity while preserving important channel-wise dependencies. The final output is obtained by rescaling the original feature maps with the activated values:

$$\tilde{x}_c = s_c \cdot x_c \tag{3.4}$$

The bottleneck reduction ratio r is set to 16 (the typical compression ratio for SE blocks), providing a balance between model complexity and performance. SE blocks are particularly effective in medical image analysis tasks such as polyp segmentation, where highlighting discriminative features is essential for accurate boundary delineation.

3.2.2 Atrous and Separable Convolutions

The proposed model incorporates atrous (dilated) convolutions drawn from DeepLabV3 [8] and separable convolutions to build more efficient encoder blocks. Atrous convolutions expand the receptive field of convolutional layers without increasing computational complexity, allowing the model to capture broader contextual information. Separable convolutions reduce computational cost by decomposing a 2D convolution into two sequential 1D convolutions to reduce the computational cost while maintaining similar performance. This allows the model to be more efficient for deployment in resource-constrained clinical environments.

3.3 Spatial Attention Gates

A key contribution of the proposed model is the integration of spatial attention gates into any U-Net backbone, following the approach of Attention U-Net [31]. Default U-Net skip connections do not explicitly highlight important spatial features, resulting in indiscriminate concatenation of encoder and decoder features, which can degrade segmentation performance.

Attention Gates (AGs) (see Figure 2.3) address this limitation by enabling the network to automatically focus on relevant structures while suppressing unimportant regions, improving segmentation accuracy particularly for smaller or irregularly shaped targets such as polyps. Each AG takes two inputs: a feature map from the encoder (skip connection) x and a gating signal from the decoder g. The attention coefficients are computed as:

$$\alpha = \sigma\left(\mathrm{Conv}_{1\times1}\left(\mathrm{ReLU}\left(\mathrm{BN}\left(\mathrm{Conv}_{1\times1}(g)\right) + \mathrm{BN}\left(\mathrm{Conv}_{1\times1}(x)\right)\right)\right)\right) \tag{3.5}$$

where x denotes the encoder skip connection features, g denotes the gating signal from the decoder, BN represents batch normalization, and σ denotes the sigmoid activation function. The computed attention coefficients $\alpha \in [0, 1]$ are directly applied to the skip connections before concatenation with the decoder features, allowing the model to adaptively weight spatial features at each stage.

3.4 Enhanced Duck Block

The DUCK-Net architecture [13] introduces a novel parallel convolution block into the U-Net encoder-decoder backbone. The original DUCK block consists of parallel convolutional branches, each performing a different operation including midscope, widescope, residual convolution, and separated convolution. The network automatically selects the most optimal branch during training to improve performance. Our proposed model enhances this parallel convolution structure by introducing two key additions: a double 7×7 separable convolution branch and an average pooling 1×1 branch. Inspired by the Inception-B block [42], the double 7×7 convolution expands the receptive field while maintaining computational efficiency. The average pooling branch preserves the original average features across the skip connections, helping to reduce information loss and better carry features from the encoder to the decoder. The updated block design is illustrated in Figure 3.1.

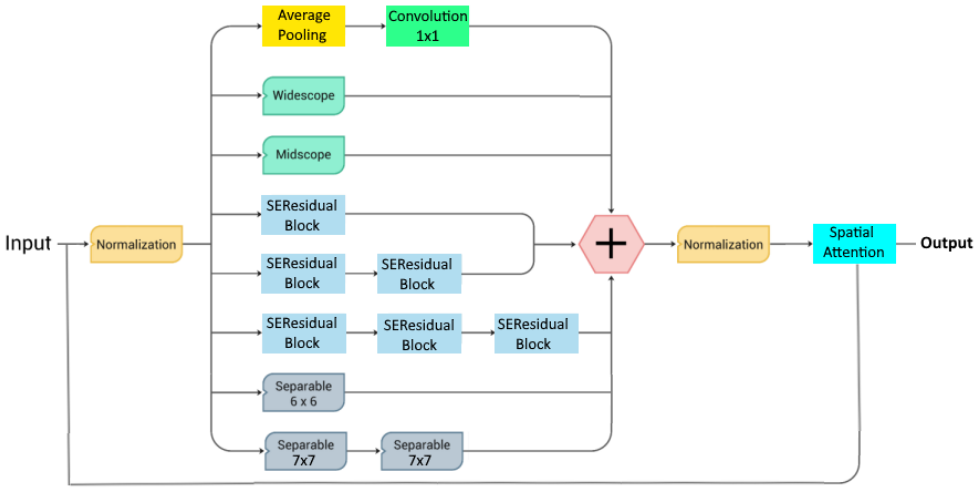

Figure 3.1: Enhanced Explainable Duck Block

3.4.1 Explainability via Attention Heatmaps

An advantage of integrating Attention Gates into the proposed architecture is that the attention coefficients can be directly visualized as heatmaps, offering an intuitive and model-specific explanation that is faithful to the segmentation decision. This approach fundamentally differs from post-hoc explanation methods such as GradCAM [38], which provide approximations that may not accurately represent the true reasoning of the model.

Moreover, attention coefficients are generated during the forward pass as part of the segmentation process; they represent a model-specific explanation that is naturally aligned with how the model arrives at its segmentation output. The spatial attention weights can be overlaid on the original input image, producing heatmaps that highlight the regions the model considers most relevant for segmentation. This makes the model's decision making process transparent, providing medical professionals with clear cues indicating areas of interest that the network focuses on.

3.4.2 Decoder and Segmentation Head

The decoder follows the standard U-Net architecture structure. The process follows progressive upsampling of feature maps to restore the original spatial resolution. At each decoder block, the upsampled features are concatenated with the corresponding attention-weighted skip connection features from the encoder. Two successive 3×3 convolutional

layers with batch normalization and ReLU activation are then applied at each stage, following the standard U-Net approach. This process continues until the feature maps are restored to the original input resolution of 320×320 pixels. A final 1×1 convolutional layer maps the feature maps to the desired output, followed by a sigmoid activation function to produce binary segmentation masks, where pixel values of 1 indicate polyp regions and 0 indicates background.

Chapter 4

Experiments

4.1 Dataset

4.1.1 Kvasir-SEG

The dataset used for all experiments is the Kvasir-SEG dataset [18], a collection of colonoscopy images annotated with ground truth segmentation masks for polyp detection. The dataset contains 1,000 polyp images, representing a variety of polyp types, sizes, and imaging conditions, making it an essential resource for training and evaluating automated polyp detection systems. The diversity of the dataset in terms of polyp appearance, size, and location encountered in clinical practice makes it a suitable benchmark for evaluating the generalizability of segmentation models. A sample of the dataset is displayed on Figure 4.1



Figure 4.1: Kvasir-SEG Dataset Sample

4.1.2 Data Preparation and Augmentation

The data preparation pipeline standardizes the input data by resizing all images and masks to 320×320 pixels. Segmentation masks are converted into binary pixel values, where 0 represents the background and 1 represents the region of interest. Image pixel values are normalized from $[0, 255]$ to $[0, 1]$ to enhance computational efficiency and accelerate model convergence. The dataset is split into a 90% training and 10% validation set, with the validation set remaining unchanged throughout all experiments to ensure unbiased evaluation.

To improve generalization and prevent overfitting, the training set undergoes a two-stage augmentation process. In the first stage, images are rotated by 90°, 180°, and 270°, expanding the dataset by a factor of 4. In the second stage, random transformations are applied, including adjustments in brightness, contrast, saturation, hue, zoom, translation, and additional rotations, introducing controlled variability while preserving the semantic integrity of the images. This augmentation strategy proved particularly effective, with results demonstrating significant improvements in model generalization across all architectures evaluated.

4.2 Evaluation Metrics

To assess segmentation performance, we employ a combination of region-based metrics alongside precision and recall. The two primary metrics used to evaluate segmentation quality are the Dice Coefficient and the Jaccard Index.

4.2.1 Dice Coefficient

The Dice Coefficient, equivalent to the F1 Score in binary segmentation, measures the overlap between the predicted segmentation mask P and the ground truth mask G:

$$\text{Dice} = \frac{2|P \cap G|}{|P| + |G|} \tag{4.1}$$

The Dice Coefficient ranges from 0 to 1, where 1 indicates perfect agreement between the predicted and ground truth masks. It is more sensitive to small segmentation errors and more forgiving to minor boundary deviations than the Jaccard Index.

4.2.2 Jaccard Index (IoU)

The Jaccard Index, also known as Intersection over Union (IoU), measures the overlap between the predicted mask P and the ground truth mask G:

$$\text{IoU} = \frac{|P \cap G|}{|P \cup G|} \tag{4.2}$$

The Jaccard Index ranges from 0, indicating no overlap, to 1, indicating a perfect match. It provides a stricter evaluation than the Dice Coefficient, as it penalizes both false positives and false negatives more heavily.

4.2.3 Precision and Recall

Precision measures the proportion of predicted positive pixels that are truly positive, while recall measures the proportion of actual positive pixels that are correctly predicted. Precision is defined as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.3}$$

where TP denotes true positives and FP denotes false positives. Recall is defined as:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4.4}$$

where FN denotes false negatives. Together with the Dice Coefficient and Jaccard Index, these metrics provide a comprehensive evaluation of segmentation quality.

4.3 Implementation Details

4.3.1 Experimental Environment

We built several U-Net architectural variants using TensorFlow from scratch to conduct experiments and compare their performance with our proposed model. Our experimental models include the original U-Net, Attention U-Net, ResU-Net, and Attention ResU-Net,

following the architectural principles described in their respective original publications. Due to computational constraints, we standardized the base filter count to 32 channels across all U-Net variants to ensure efficient training while maintaining model capacity.

4.3.2 Optimiser and Hyperparameters

All models were trained using consistent hyperparameters to ensure fair comparisons. The Adam optimizer [22] was employed with an initial learning rate of 1×10^{-4} and a weight decay rate of 1×10^{-8}, using $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Models were trained for 50 epochs, achieving convergence within approximately 12 hours of GPU time. To prevent overfitting, dropout layers with a rate of 0.25 were incorporated throughout all network architectures. A batch size of 16 was selected to balance model performance with the hardware memory constraints.

4.3.3 Loss Function

All models were trained using the Dice Coefficient loss, defined as:

$$\mathcal{L}_{\text{Dice}} = 1 - \text{Dice} \tag{4.5}$$

This loss function directly optimizes the Dice Coefficient, making it particularly suitable for binary segmentation tasks where class imbalance between foreground and background regions is common, as is the case in polyp segmentation. We utilize this loss function as our goal is to maximize our dice score.

4.3.4 Model Configurations

Several U-Net architectural variants were implemented and evaluated alongside two state-of-the-art architectures. For DUCK-Net, computational constraints required the use of the 17-filter variant rather than the original 34-filter architecture, as the larger model exceeded the available GPU memory.

Chapter 5

Results and Discussion

5.1 Quantitative Results

Table 5.1 presents the performance metrics of all evaluated models on the Kvasir-SEG dataset, comparing the Dice Coefficient, Jaccard Index, Precision, Recall, and Accuracy, as well as whether each model supports explainability through attention-based heatmaps.

Table 5.1: Architecture comparisons and model support of explanation on Kvasir-SEG

Model	**Dice**	**Jaccard**	**Precision**	**Recall**	**Accuracy**	**XAI**
Base U-Net	0.8005	0.7972	0.8717	0.7934	0.9363	N
Attention U-Net	0.7943	0.7913	0.8606	0.8180	0.9318	Y
ResU-Net	0.7960	0.7918	0.8623	0.7981	0.9335	N
Attention ResU-Net	0.7934	0.7873	0.8886	0.8165	0.9319	Y
DUCK-Net (17 filters)	0.8725	0.8643	0.8951	0.8937	0.9598	N
Proposed Model (17 filters)	0.8621	0.8575	0.8871	0.8892	0.9559	Y

The base U-Net achieved a Dice score of 0.8005 and a Jaccard Index of 0.7972, performing well above the typical U-Net IoU baseline of approximately 0.76 reported in the literature. This improvement can be largely attributed to the two-stage data augmentation strategy employed, which significantly improved model generalization by expanding the training dataset and introducing controlled variability.

Among the attention-based U-Net variants, the Attention U-Net and Attention ResU-Net achieved Dice scores of 0.7943 and 0.7934 respectively, performing slightly below the

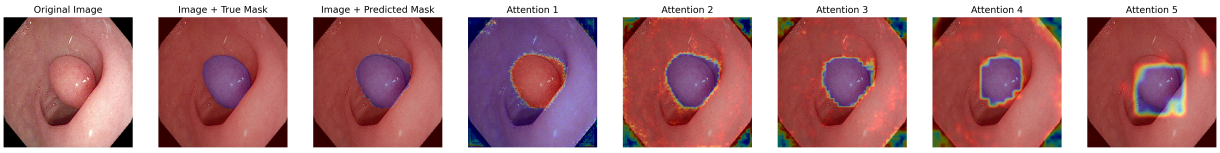

Figure 5.1: Multi-scale Attention Heatmaps

base U-Net. This somewhat unexpected result is likely attributable to the limited number of training epochs, as both models showed continuing improvements in precision and recall during the final training stages, suggesting that 50 epochs were insufficient for full convergence of the attention mechanisms.

DUCK-Net achieved the highest overall Dice score of 0.8725 among all evaluated models. This indicates that the parallel processing blocks introduced by DUCK-Net are particularly effective for this segmentation task.

The proposed model achieved a Dice score of 0.8621 and a Jaccard Index of 0.8575, placing it as the best performing explainable model among all evaluated architectures. While it did not surpass DUCK-Net in terms of raw segmentation performance, it maintained highly competitive results while uniquely supporting model-specific explainability through attention heatmaps. This represents a meaningful contribution, as DUCK-Net provides no mechanism for generating visual explanations of its segmentation decisions.

5.2 Qualitative Results and Attention Heatmap Visualisation

Figure 5.1 presents the predicted segmentation masks alongside the attention heatmaps generated by the proposed model for a sample polyp image from the Kvasir-SEG dataset. The heatmaps are generated directly from the spatial attention weights at each stage of the network, providing a model-specific visual explanation of the segmentation decision.

The figure presents a predicted mask alongside attention heatmaps, where the areas the model is focusing on appear in blue, while regions it is not attending to are shown in red. The first attention map appears to focus on regions where the polyp is absent, while attentions 2 to 5 highlight the polyp at different scales. The attention heatmaps reveal a structured and interpretable decision-making process. Our model first distinguishes between areas where the polyp is absent and where it is present, effectively separating the two before making its final prediction. The initial attention map appears to suppress

regions that do not contain the polyp, helping the model focus on relevant structures. As the attention progresses through different levels, the model hones in on the polyp with increasing precision, capturing both fine details and broader contextual information. This layer-by-layer progression demonstrates that the model systematically refines its prediction rather than making an arbitrary segmentation decision.

This structured approach to segmentation is particularly valuable in medical imaging, where understanding how a model arrives at its output is as important as the output itself. By visualizing how different attention layers contribute to the final segmentation, clinicians can assess whether the model is focusing on clinically relevant anatomical structures or being misled by irrelevant features. These insights can also help identify potential failure cases, guide future model improvements, and support the responsible integration of AI into clinical workflows.

Chapter 6

Discussion

6.1 Interpretation of Results

The results demonstrate that integrating attention mechanisms into a U-Net based architecture allows for competitive segmentation performance while simultaneously supporting model-specific explainability. The proposed model achieved the best performance among all explainable models evaluated, outperforming the attention-based U-Net variants.

A key observation from the experiments is the significant impact of data augmentation on model performance. The two-stage augmentation strategy substantially improved generalization across all evaluated architectures, with the base U-Net achieving results well above what is typically reported in the literature. This underscores the importance of data augmentation for medical image segmentation tasks, where datasets are often limited in size due to the challenges of acquiring and annotating medical images.

The attention-based U-Net variants, despite incorporating sophisticated mechanisms, performed similarly to or slightly below the base U-Net. This can be attributed to the limited training duration of 50 epochs, as these models showed continued improvement during the final training stages. This observation suggests that attention mechanisms may require extended training periods to fully develop their pattern recognition capabilities, and that further performance gains are achievable with additional training.

6.2 Comparison with Post-hoc Explainability Methods

A fundamental advantage of the proposed model-specific approach is that the attention heatmaps are generated as an inherent part of the network's forward pass, rather than being approximated after the fact. Unlike post-hoc methods such as GradCAM, which compute gradient-based saliency maps after inference, the attention weights in the proposed model directly reflect the network's actual decision-making process at each layer. This fine-grained, layer-by-layer visualization is crucial for medical imaging tasks where trust, interpretability, and clinical validation are essential requirements.

6.3 Limitations

While the proposed model demonstrates promising results, several limitations should be noted. First, the model was evaluated on a single dataset, the Kvasir-SEG, and further validation on additional polyp segmentation datasets would be necessary to demonstrate cross-dataset generalizability. Second, the training duration of 50 epochs was insufficient for full convergence of the attention-based models, and hyperparameter tuning tailored to each individual architecture would likely yield further performance improvements. Third, applying a uniform learning rate and batch size across all models may not fully reflect their individual performance potential, as different architectures may benefit from different training configurations.

These limitations motivate the next stage of this research, presented in Part Two, where a more powerful dual encoder architecture is explored that combines the complementary strengths of convolutional neural networks and vision transformers to achieve improved segmentation performance and richer multi-scale explanations.

Part II

Dual Encoder Explainable Medical Image Segmentation via Attention-Gated Fusion

*An earlier version of this work will appear in [21]. Code available at:
https://github.com/MaudDK/MedSeg-XAI-AGFusion

Chapter 7

Methodology

7.1 Model Overview

This chapter presents the proposed dual encoder architecture (see Figure 7.1 for explainable medical image segmentation.

Building upon Part One, which portrayed the value of model-specific explainability through attention mechanisms in CNN-based architectures, this work explores whether combining the strengths of convolutional neural networks and vision transformers can yield improved segmentation performance alongside richer multi-scale visual explanations.

The proposed architecture employs two parallel encoders: a pretrained ResNet-34 CNN encoder to capture fine-grained local spatial features, and a pretrained Swin Transformer encoder to capture global contextual information across the entire image. The two encoders are fused at multiple scales using a learned attention-gated fusion mechanism, which selectively combines the outputs of both encoders based on their relevance to the segmentation task.

As a byproduct of this fusion process, the architecture generates multi-scale attention heatmaps that serve as model-specific visual explanations, highlighting the regions that contributed most to the segmentation decision.

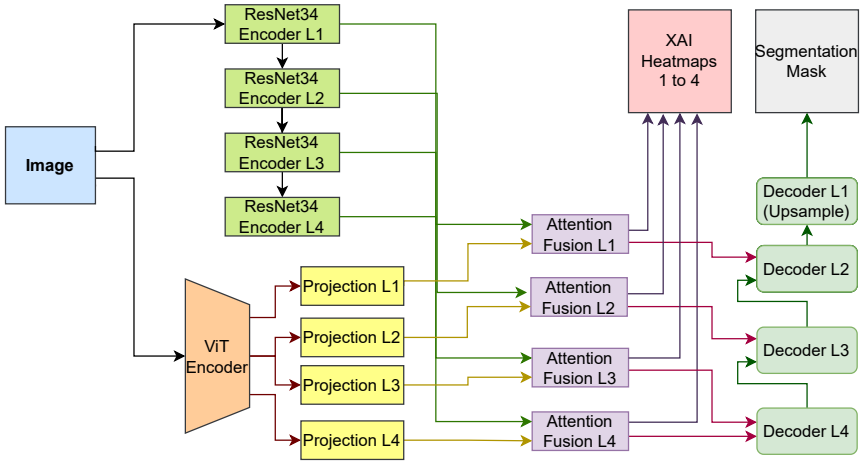

Figure 7.1: Dual Encoder

7.2 Spatial Feature Encoding

7.2.1 ResNet34 Encoder

The ResNet-34 encoder serves as the CNN backbone of the dual encoder architecture, responsible for capturing fine-grained local spatial features through hierarchical convolutional operations.

ResNet-34 integrates residual learning principles into its architecture, incorporating skip connections within each encoder block to facilitate gradient flow during training and enable the construction of deeper networks while mitigating the vanishing gradient problem [16].

The architecture processes input images at a resolution of 384×384 pixels, producing hierarchical feature maps at four scales: 96×96 (128 channels), 48×48 (128 channels), 24×24 (256 channels), and 12×12 (512 channels). These multi-scale feature maps capture both fine-grained spatial details at higher resolutions and more abstract semantic information at lower resolutions, providing a rich representation of local image features for the fusion mechanism.

By leveraging pretrained weights from ImageNet, the ResNet-34 encoder significantly reduces the computational requirements of training and enables the model to perform effectively without the large dataset requirements typically associated with training deep networks from scratch.

7.3 Global Feature Encoding

7.3.1 Swin Transformer-Base Encoder

The Swin Transformer encoder serves as the vision transformer backbone of the dual encoder architecture, responsible for capturing global contextual information across the entire image.

Unlike CNN-based encoders, which are inherently limited by their local receptive fields imposed by fixed-size convolutional kernels, the Swin Transformer leverages self-attention mechanisms that allow each image patch to attend to all other patches, enabling the capture of long-range spatial relationships and global context [27].

The importance of global context in medical image segmentation is particularly evident for tasks such as polyp detection, where segmentation targets may vary significantly in size, appearance, and spatial location across the image. Understanding relationships between distant yet related regions is essential for accurate delineation of such targets, making the global contextual modeling of the Swin Transformer a valuable complement to the local feature extraction of the ResNet-34 encoder.

Specifically, the pretrained `swin_base_patch4_window12_384` model is employed as the transformer encoder, processing input images at 384×384 resolution. The Swin Transformer produces hierarchical feature maps at four corresponding scales with higher channel dimensions than the ResNet-34 encoder: 96×96 (128 channels), 48×48 (256 channels), 24×24 (512 channels), and 12×12 (1024 channels).

7.3.2 Channel Projection

Transformer feature maps must first be aligned with those of the ResNet-34 encoder. This is achieved by applying 1×1 convolutional projection layers to reduce the Swin Transformer features to match the ResNet-34 channel dimensions at each scale. This lightweight projection preserves the semantic content of the transformer features while enabling direct comparison and fusion with the CNN features.

7.4 Dual Encoder Attention Fusion

7.4.1 Attention Gate Mechanism

The proposed architecture fuses features from both encoders at each skip connection level through a modified attention gating mechanism. The mechanism learns to selectively combine the CNN and transformer features based on their relevance to the segmentation task at each spatial location, producing a fused feature representation that leverages both local and global information. The attention gate takes as input the ResNet-34 skip connection features x and the projected Swin Transformer features g, and computes the attention map as:

$$\psi = \sigma\left(\text{Conv}_{1\times1}\left(\text{ReLU}\left(\text{BN}\left(\text{Conv}_{1\times1}(g)\right) + \text{BN}\left(\text{Conv}_{1\times1}(x)\right)\right)\right)\right) \tag{7.1}$$

where x denotes the ResNet-34 skip connection features, g denotes the Swin Transformer features serving as the gating signal, BN represents batch normalization, and σ denotes the sigmoid activation function. The learned attention map $\psi \in [0, 1]$ determines the contribution of each encoder at every spatial location, and the final fused feature representation is computed as:

$$f_{\text{fused}} = \psi \cdot x + (1 - \psi) \cdot g \tag{7.2}$$

When attention values approach 1, the fused features rely more heavily on the ResNet-34 encoder's local spatial details. When attention values approach 0, the Swin Transformer encoder's global contextual information dominates. This weighting is learned end-to-end during training and varies spatially across the image, allowing the model to adaptively leverage local and global information where each is most needed.

7.4.2 Swin-Guided ResNet Attention

The Swin Transformer features are used as the gating signal g in the attention mechanism because the global understanding of the image provided by the transformer encoder offers high-level semantic guidance for determining which local CNN features are most relevant. This design aligns with the principle that global context should inform local feature selection: the Swin Transformer's holistic view of anatomical structures identifies which fine-grained details captured by the ResNet-34 encoder warrant emphasis in the final segmentation. Ablation experiments presented in Chapter 9 confirm that Swin-guided

attention consistently outperforms both ResNet-guided and dual-guided attention configurations, achieving the highest Dice score of 0.907 among all evaluated attention strategies.

7.4.3 Multi-Scale Fusion

The attention-gated fusion is applied at all four skip connection levels of the architecture, producing fused feature representations at scales of 12×12, 24×24, 48×48, and 96×96. This multi-scale fusion enables the model to combine local and global information at multiple levels of abstraction, capturing both fine-grained boundary details and broader semantic context simultaneously.

7.5 Decoder and Segmentation Head

7.5.1 Decoder Blocks

The decoder follows a standard U-Net structure with skip connections. At each decoder block, the feature maps are upsampled by a factor of 2 using bilinear interpolation, concatenated with the corresponding fused features from the attention-gated skip connections, and passed through two successive 3×3 convolutional layers with batch normalization and ReLU activation. The decoder progressively upsamples from a resolution of 12×12 to 96×96, restoring spatial detail at each stage.

7.5.2 Final Upsampling

Following the final decoder block, an additional upsampling module increases the resolution from 96×96 to the original input resolution of 384×384 pixels using bilinear interpolation. This ensures the output segmentation mask is aligned with the original input image, enabling pixel-precise segmentation predictions.

7.5.3 Segmentation Output

A final 1×1 convolutional layer maps the upsampled feature maps to the desired output, followed by a sigmoid activation function to produce the binary segmentation mask.

Pixel values of 1 indicate polyp regions and pixel values of 0 indicate background, providing a pixel-wise prediction that can be directly overlaid on the original colonoscopy image for clinical interpretation.

7.6 Explainability via Attention Weights

7.6.1 Attention Weight Extraction

A key advantage of the attention-gated fusion mechanism is that it generates model-intrinsic visual explanations as a natural byproduct of the fusion process. The attention maps ψ produced at each encoder level directly represent the model's decision-making process, showing which encoder features contributed to the segmentation at different spatial locations and scales. This fundamentally differs from the model-specific approach presented in Part One, where attention weights from skip connections within a single CNN encoder were visualized.

In this work, the attention maps represent the relative contribution of two complementary encoders at each spatial location, providing a richer and more informative explanation of the segmentation decision. Furthermore, since the attention maps are generated during the forward pass as part of the fusion process, they provide a more faithful explanation of the model's reasoning than post-hoc methods such as GradCAM, which approximate saliency after inference.

7.6.2 Visualization

The attention maps generated at each of the four fusion levels are visualized as heatmaps and overlaid on the original input image, providing clinicians with interpretable visual cues that highlight the regions most relevant to the segmentation decision at different scales.

As shown in the results presented in Chapter 9, the attention maps reveal a clear hierarchical focus: the first two maps primarily highlight the edges of the polyp, capturing boundary information, while the subsequent maps concentrate on the polyp region itself, guiding the network towards the final segmentation.

This progression demonstrates how the model sequentially refines its spatial focus from edges to the full object, providing clinicians with an interpretable window into the model's decision-making process at each stage.

Chapter 8

Experiments

8.1 Data

8.1.1 Kvasir-SEG

The dataset used for all experiments is the Kvasir-SEG dataset [18], the same benchmark used in Part One, enabling direct comparison of results across both parts of this thesis.

The dataset contains 1,000 colonoscopy images annotated with pixel-precise ground truth segmentation masks for polyp detection, capturing the variability in polyp appearance, size, and location encountered in clinical practice.For this work, all images and masks are resized to 384×384 pixels, a higher resolution than the 320×320 used in Part One, to accommodate the input requirements of the Swin Transformer encoder.

8.1.2 Data Augmentation

To improve generalization and prevent overfitting, the training set undergoes the same two-stage augmentation strategy employed in Part One. In the first stage, images are rotated by 90°, 180°, and 270°, expanding the dataset by a factor of 4. In the second stage, random transformations are applied, including adjustments in brightness, contrast, saturation, hue, zoom, translation, and additional rotations, introducing controlled variability while preserving the semantic integrity of the images. As demonstrated in Part One, this augmentation strategy proved particularly effective for improving model generalization, and is retained here to ensure consistency across experiments.

8.2 Metrics

8.2.1 Dice Coefficient

As in Part One, the Dice Coefficient is used as the primary metric for evaluating segmentation performance. The Dice Coefficient measures the overlap between the predicted segmentation mask P and the ground truth mask G:

$$\text{Dice} = \frac{2|P \cap G|}{|P| + |G|} \tag{8.1}$$

8.2.2 Jaccard Index

The Jaccard Index, or Intersection over Union (IoU), provides a stricter evaluation of segmentation performance:

$$\text{IoU} = \frac{|P \cap G|}{|P \cup G|} \tag{8.2}$$

Together with the Dice Coefficient, the Jaccard Index provides a comprehensive evaluation of segmentation quality, penalizing both false positives and false negatives.

8.2.3 Binary Cross Entropy Loss

In addition to the Dice-based metrics, Binary Cross Entropy (BCE) loss is employed during training alongside the Dice loss to provide pixel-wise supervision.

BCE loss is defined as:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right] \tag{8.3}$$

where y_i denotes the ground truth label and $\hat{y}_i$ denotes the predicted probability for pixel i.

8.3 Implementation Details

8.3.1 Experiment Environment

All experiments were performed using PyTorch, in contrast to the TensorFlow implementation used in Part One. The model was trained on a single GPU, leveraging pretrained weights for both the ResNet-34 and Swin Transformer encoders to reduce computational requirements and enable effective fine-tuning on the Kvasir-SEG dataset.

8.3.2 Loss Function

The model is trained using a combined loss function that sums the Dice Coefficient loss and Binary Cross Entropy loss:

$$\mathcal{L} = \mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{BCE}} \tag{8.4}$$

This combined loss provides both region-level supervision through the Dice loss and pixel-level supervision through the BCE loss, encouraging the model to produce accurate segmentation boundaries as well as correct region predictions.

8.3.3 Optimiser and Learning Rate

The model is trained using the Adam optimizer [22] with a learning rate scheduler to improve convergence.

Pretrained weights are leveraged for both encoders, with fine-tuning applied to adapt the pretrained representations to the medical image segmentation task. Parameter freezing configurations are explored as part of the ablation study, with results presented in Chapter 9.

8.3.4 Model Configuration

The architecture processes input images at 384×384 resolution. The ResNet-34 encoder produces feature maps at four scales, and the Swin Transformer encoder produces corresponding feature maps with higher channel dimensions, as described in Chapter 7. Channel projection layers reduce the Swin Transformer feature dimensions to match those of the

ResNet-34 encoder at each scale prior to attention-gated fusion. Several model configurations are evaluated as part of the ablation study, including variations in attention guidance strategy and parameter freezing, to determine the optimal configuration for the dual encoder architecture.

Chapter 9

Results and Discussion

9.1 Model Performance

9.1.1 Quantitative Results

Table 9.1 presents a comprehensive comparison of training strategies evaluated during the development of the dual encoder architecture, examining the impact of training from scratch, data augmentation, and pretrained weights on segmentation performance across all encoder architectures.

Table 9.1: Performance comparison across different training strategies

Model	From Scratch				Augmented				Aug & Pretrained			
	Dice	**IoU**	**Prec.**	**Rec.**	**Dice**	**IoU**	**Prec.**	**Rec.**	**Dice**	**IoU**	**Prec.**	**Rec.**
ViT-Tiny	0.603	0.480	0.718	0.636	0.724	0.618	0.787	0.782	**0.872**	**0.809**	**0.869**	**0.915**
ViT-Small	0.645	0.527	0.709	0.709	0.697	0.592	0.743	0.759	**0.880**	**0.810**	**0.901**	**0.899**
ViT-Base	0.607	0.480	0.670	0.679	0.674	0.558	0.669	0.801	**0.875**	**0.807**	**0.890**	**0.905**
ResNet-34	0.854	0.788	0.890	0.874	0.882	0.817	0.894	0.910	**0.893**	**0.837**	**0.910**	**0.910**

The results demonstrate a clear and consistent trend across all architectures: training from scratch yields the lowest performance, augmentation alone provides a meaningful improvement, and combining augmentation with pretrained weights yields the greatest performance gains across all models. This finding is consistent with the results observed in Part One, further reinforcing the importance of data augmentation for medical image segmentation tasks where datasets are inherently limited in size.

A notable observation is that ViT models trained from scratch struggle significantly to generalize, achieving Dice scores between 0.603 and 0.645, while the ResNet-34 model adapts more quickly, achieving a Dice score of 0.854 from scratch. This reflects the well-known data hunger of vision transformer models, which require substantially more training data and pretraining to generalize effectively compared to CNN-based models. When pretrained weights are combined with augmentation however, ViT models achieve performance comparable to the ResNet-34 encoder, with Dice scores ranging from 0.872 to 0.880, confirming that pretraining is essential for effective transformer deployment in medical imaging contexts where data is scarce.

9.1.2 Qualitative Results

Figure 9.1 presents the predicted segmentation masks alongside the multi-scale attention heatmaps generated by the proposed dual encoder architecture for a sample polyp image from the Kvasir-SEG dataset. The attention maps reveal a clear hierarchical focus across the four fusion levels. The first two attention maps primarily highlight the edges of the polyp, capturing boundary information, while the subsequent maps concentrate on the polyp region itself, guiding the network towards the final segmentation. This progression demonstrates how the model sequentially refines its spatial focus from boundary details to the full object, providing clinicians with an interpretable window into the model's decision-making process at each scale.

9.2 Ablation Study

9.2.1 Skip Connections

Table 9.2 presents the results of the skip connection ablation study, evaluating the impact of skip connections on segmentation performance across all encoder architectures.

Skip connections consistently improve segmentation performance across all encoder architectures, with Dice improvements ranging from 0.9% to 1.1%. These results validate the importance of skip connections in preserving fine-grained spatial information and enhancing feature propagation from encoder to decoder, consistent with the original U-Net design principles. Based on these results, the Swin Transformer is selected as the transformer-based encoder for the dual encoder architecture due to its superior performance with skip connections, achieving a Dice score of 0.899. The ResNet-34 encoder is selected as the CNN-based encoder for its competitive performance and more efficient inference times.

Table 9.2: Performance comparison of pretrained models with and without skip connections

Model	Skip Connections	Dice	IoU	Precision	Recall
ResNet-34 UNet	With Skip	**0.893**	**0.837**	0.910	**0.910**
	No Skip	0.883	0.819	**0.925**	0.886
DeepLabV3+	With Skip	**0.890**	**0.825**	**0.890**	0.923
	No Skip	0.880	0.814	0.881	**0.923**
Swin Transformer	With Skip	**0.899**	**0.840**	0.897	**0.935**
	No Skip	0.890	0.829	**0.901**	0.920

9.2.2 Attention Guidance and Parameter Freezing

Table 9.3 presents the results of the attention guidance ablation study, evaluating the impact of different attention configurations and parameter freezing strategies on the dual encoder architecture.

Table 9.3: Performance comparison of Attention Guidance & Parameter Freezing in dual encoder architectures

Model	Attention Guidance	Frozen	Dice	IoU	Precision	Recall
No Attention Dual Encoder	None	Unfrozen	0.902	0.844	0.916	0.922
Attention Dual Encoder	Swin-guided	Frozen	**0.907**	**0.855**	**0.926**	0.921
	Res34-guided	Frozen	0.904	0.850	0.921	0.919
Double Attention Dual Encoder	Dual-guided	Frozen	**0.906**	**0.853**	**0.922**	0.923
	Dual-guided	Unfrozen	0.901	0.846	0.907	**0.934**
Double Weighted Attention Dual Encoder	Dual-guided	Frozen	**0.903**	**0.848**	**0.906**	0.934
	Dual-guided	Unfrozen	0.902	0.841	0.894	**0.941**
Weighted Double Dual Encoder Attention	Dual-guided	Unfrozen	0.900	0.846	0.927	0.909

Incorporating attention mechanisms provides consistent improvements over the no-attention baseline across all evaluated configurations. The Swin-guided Attention Dual Encoder achieves the highest Dice score of 0.907 and IoU of 0.855, demonstrating that selective feature emphasis from the transformer pathway is more effective than bidirectional attention configurations. Single attention guidance from the Swin encoder outperforms dual-guided configurations, suggesting that using the transformer's global understanding to guide local CNN feature selection is a more effective strategy than allowing both encoders to guide each other simultaneously. Parameter freezing in dual-guided architectures maintains competitive performance while reducing computational overhead, offering a practical trade-off for resource-constrained deployment scenarios.

9.3 Comparison with State of the Art

Table 9.4 compares the best dual encoder configuration against state-of-the-art transformer and CNN-based segmentation models on the Kvasir-SEG dataset.

Table 9.4: Performance comparison of best models across different architectures

Model	Dice	IoU	Precision	Recall
Dual Encoder Architectures				
Attention Dual Encoder (Swin-guided)	**0.907**	**0.855**	**0.926**	0.921
Transformer Architectures				
Swin Base (Pretrained, Skip)	0.899	0.840	0.897	0.935
ViT-Small (Pretrained)	0.880	0.810	0.901	0.899
ViT-Base (Pretrained)	0.875	0.807	0.890	0.905
ViT-Tiny (Pretrained)	0.872	0.809	0.869	0.915
CNN Architectures				
ResNet-34 UNet (Pretrained)	0.893	0.837	0.910	0.910
DuckNet (Augmented)	0.889	0.824	0.898	0.917
DuckNet-34 (Augmented)	0.881	0.816	0.900	0.905

The Swin-guided Attention Dual Encoder achieves the highest overall performance across all evaluated architectures, with a Dice score of 0.907 and an IoU of 0.855. This represents a meaningful improvement over the best individual encoder architectures, with the dual encoder outperforming the Swin Transformer alone by 0.8% in Dice and the ResNet-34 UNet by 1.4% in Dice. These results demonstrate that the attention-gated fusion mechanism effectively combines the complementary strengths of the transformer and CNN encoders, achieving superior segmentation performance compared to either encoder in isolation.

It is also worth noting that the proposed dual encoder architecture achieves this improved performance while simultaneously generating multi-scale explainability heatmaps displayed in 9.1, a capability that none of the single encoder baselines provide. This positions the proposed architecture as both the highest performing and most interpretable model evaluated, directly addressing the core motivation of this thesis: achieving accurate and explainable medical image segmentation.

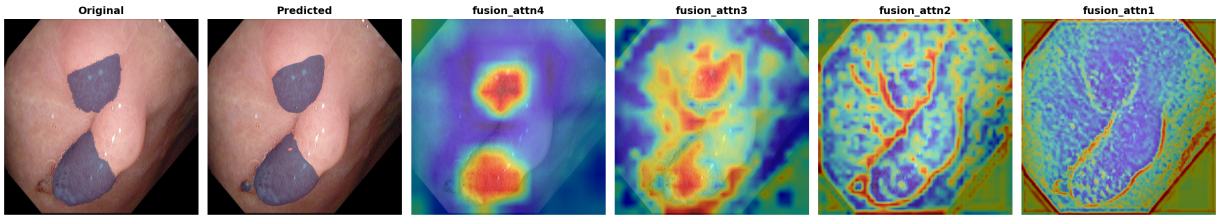

Figure 9.1: Dual Encoder Attention Heatmaps

9.4 Explainability Analysis

9.4.1 Attention Weight Visualisation

The multi-scale attention maps generated by the proposed architecture provide a hierarchical and interpretable view of the model's decision-making process. As illustrated in Figure 9.1, the attention maps at the four fusion levels reveal a clear progression from boundary detection to full region segmentation. The first two attention maps primarily focus on the edges of the polyp, capturing fine-grained boundary information at higher resolutions. The subsequent attention maps shift their focus towards the polyp region as a whole, providing broader contextual guidance for the final segmentation decision. This structured progression demonstrates that the model systematically refines its prediction across scales, rather than making an arbitrary decision, providing clinicians with a transparent and interpretable view of how the segmentation was produced. The attention maps also serve as a diagnostic tool: if the model segments a region it is not attending to, this inconsistency may flag a potentially unsafe and unreliable prediction, helping clinicians screen multiple patients efficiently, identify failure cases, and build justified trust in the model's output.

9.4.2 Comparison with Post-hoc Methods

The attention maps generated by the proposed architecture are fundamentally different from post-hoc explanation methods such as GradCAM in two important ways. First, the attention maps are generated during the forward pass as an inherent part of the fusion mechanism, rather than being approximated after inference. This ensures that the explanations are faithful to the model's actual decision-making process, rather than being a post-hoc approximation. Second, the multi-scale nature of the attention maps provides a richer and more informative explanation than a single saliency map, offering clinicians insight into how the model's focus evolves across different levels of abstraction. Together,

these properties make the proposed architecture's explainability mechanism more clinically valuable than model-agnostic post-hoc methods, supporting greater trust and transparency in AI-assisted clinical decision making.

9.5 Discussion

9.5.1 Interpretation of Results

The results presented in this chapter demonstrate that the proposed dual encoder architecture achieves state-of-the-art segmentation performance on the Kvasir-SEG dataset while simultaneously generating faithful, multi-scale visual explanations. The ablation study confirms that each component of the architecture contributes meaningfully to the overall performance: pretrained weights are essential for effective transformer deployment, skip connections consistently improve spatial feature propagation, and Swin-guided attention provides the most effective fusion strategy.

The consistent superiority of the Swin-guided attention configuration over dual-guided and ResNet-guided alternatives provides important insight into the complementary nature of the two encoders. The global semantic understanding provided by the Swin Transformer is well-suited to guide the selection of locally relevant features from the ResNet-34 encoder, while the reverse is less effective, suggesting an asymmetry in the informativeness of the two encoders' representations for this task.

9.5.2 Limitations

While the proposed architecture demonstrates strong performance, several limitations should be acknowledged. The model was evaluated on a single dataset, the Kvasir-SEG, and further validation on additional medical imaging datasets and modalities would be necessary to demonstrate the generalizability of the approach. The computational requirements of the dual encoder architecture are higher than those of single encoder models, which may present challenges for deployment in resource-constrained clinical environments. Additionally, while the individual attention maps at each fusion level provide interpretable visualizations, combining these multi-scale maps into a single composite explanation could provide a more clinically actionable visualization, and represents a promising direction for future work. These limitations and directions for future research are discussed further in the closing chapters of this thesis.

Part III

Contributions of Thesis and Future Work Towards Multimodal Explainable AI

Chapter 10

Reflections and Contributions

10.1 Summary of Contributions

This thesis presented two complementary frameworks for explainable medical image segmentation. Building upon each innovation of the previous framework, we advance the state of the art in both segmentation accuracy and the quality of model-specific visual explanations provided to medical professionals.

Part One introduced an attention-based CNN architecture for explainable medical image segmentation. The architecture demonstrates that it is possible to achieve competitive segmentation performance while simultaneously generating faithful, model-specific explanations by embedding spatial attention gates directly into the network architecture. The key contribution of Part One includes model-specific explanations generated as part of the forward pass of the model. These explanations are more faithful to the model's decision-making process than model-agnostic post-hoc methods such as GradCAM. The proposed architecture achieved the best performance, a 0.8621 Dice score across all explainable models on the Kvasir-SEG dataset. In addition, the model supported interpretable heatmaps that are faithful and provide clinicians with a transparent view of how the segmentation was produced.

Part Two advances the previous framework by introducing an explainable dual encoder architecture that fuses a pretrained ResNet-34 CNN encoder with a pretrained Swin Transformer encoder. Moreover, the dual encoder fuses both encoder representations through a custom learned attention-gated mechanism at multiple scales. The main contribution of Part Two is bridging a limitation of CNN encoders: their reliance on local spatial feature extraction without the ability to model global context. Part Two's model combines

the local spatial feature extraction of CNNs with the global contextual modeling of vision transformers, effectively fusing them through an attention-gated fusion mechanism that automatically weights each model's contributions. In turn, yielding state-of-the-art segmentation performance compared to single encoder mechanisms, while also simultaneously generating rich multi-scale visual explanations. The Swin-guided Attention Dual Encoder achieved the best overall performance across all architectures, with a Dice score of 0.907 and an IoU of 0.855, outperforming all single encoder baselines.

These two contributions represents an impactful step forward for explainable AI in medical imaging, demonstrating that accuracy and interpretability can co-exist and be generated simultaneously through innovative architecture design.

10.2 Reflection on Model-Specific Explainability

A recurrent theme of this thesis is the argument for model-specific explainability over model-agnostic post-hoc methods. We insist that faithfulness of explanations to the model's actual reasoning is essential for clinical trust and model validation. In contrast, post-hoc methods give an approximation of the model's reasoning. This distinction is particularly important in medical imaging.

Model-agnostic post-hoc methods such as GradCAM, LIME, and SHAP generate explanations after inference. This results in multiple passes reducing computational efficiency. Additionally, these post-hoc methods approximate the model's behavior from the outside. These methods are widely adopted due to their simplicity and ability to be applied to any model. However, they are decoupled from the model's internal decision making process and produce explanations that may not reflect how the output is decided. The connotations of this limitation are concerning in clinical settings, where an explanation that does not faithfully reflect the model's decision may mislead medical professionals.

In contrast, the attention-based mechanisms in both Part One and Part Two of this thesis generate the explanations as an inherent part of the model's forward pass. This process directly reflects the model's actual decision-making process. The attention coefficients computed during the fusion represent the model's ability to weigh which spatial locations are most significant for the segmentation task. Additionally, the attention coefficients provide faithful and interpretable visualization of the model's decision-making process.

The evolution from Part One to Part Two also demonstrates how model-specific explainability can be enhanced alongside improvements in segmentation accuracy. Part One

generated attention heatmaps from a single CNN encoder. Part Two further evolves and generates multi-scale attention heatmaps that represent the combined contribution of both the CNN and Transformer encoders, providing a richer and more informative explanation of the segmentation decision. Additionally, it introduces a novel method to fuse two completely different architectures into a single cohesive model, demonstrating that complementary encoder representations can be combined in a way that improves both segmentation and explainability simultaneously.

10.3 Relationship to Prior Work

The frameworks proposed in this thesis build upon and advances prior work in medical image segmentation and explainable AI.

With respect to CNN-based segmentation, the work in Part One builds directly upon the U-Net [35] and Attention U-Net [31] architectures, extending the attention gate formulation to support model-specific explainability with a more powerful architectural backbone inspired by DUCK-Net [13]. The proposed architecture achieves competitive performance with DUCK-Net while uniquely supporting explainability. In addition, it addresses a key limitation of the state-of-the-art CNN-based approach.

In terms of hybrid CNN-transformer segmentation, the work in Part Two improves upon TransFuse [51]. Both approaches employ a parallel dual encoder design with CNN and transformer encoders fused at multiple scales. However, the proposed architecture differs in two important respects. First, the transformer component is a Swin Transformer-Base rather than a standard ViT. Second, the fusion employs an attention gate in which the Swin Transformer features explicitly guides what spatial regions to attend to within the ResNet-34 feature maps; in constrast, TransFuse uses additive fusion. This attention-gated fusion not only improves segmentation performance but also generates model-specific visual explanations, which TransFuse does not support.

10.4 Looking Ahead

The contributions of this thesis presents a strong foundation for several important directions for future work. The visual explanations produced by the dual encoder architecture in Part Two represent a meaningful step forward, however, visual heatmaps alone may be insufficient for medical professionals who may require additional contextual understanding. A clinician examining a segmentation of a suspicious polyp may benefit not only from

knowing where the model focused its attention, but also from a language description of the polyp's characteristics, potential cancer risk, and extra context.

This motivates the extensions discussed in future work, which outlines a method for integrating textual explanations alongside visual ones. Additionally, future work in applying the framework across additional medical imaging domains including echocardiogram analysis for ejection fraction computation and breast cancer segmentation. These extensions represent the next steps in the broader research, towards a fully multimodal and clinically deployable explainable AI system for medical image segmentation.

Chapter 11

Future Work and Conclusion

11.1 Multimodal Textual Explanation via RAG and LLM

Visual heatmaps communicate where the model focused its attention; however they do not explain what was identified or why it is clinically significant. Medical professionals may require a natural language description of the segmentation target, including characteristics such as size, morphology, and potential malignancy risk.

We propose a pipeline that leverages the Swin Transformer encoder embeddings from the dual encoder architecture as input to SigLIP, a vision-language model that aligns image embeddings with text caption queries. The queries generated by SigLIP are passed to a Retrieval Augmented Generation (RAG) system, which retrieves relevant medical context from a curated knowledge base of clinical literature specific to the identified anomaly. The retrieved context is then passed to a Large Language Model (LLM), which generates a natural language description of the segmentation target and its clinical implications.

This combined system provides two complementary accounts of the model's reasoning. The attention heatmaps show where the model focused, while the LLM-generated text explains what was found and why it matters clinically. The overall pipeline is as follows:

1. The input medical image is processed by the dual encoder architecture, producing a segmentation mask and multi-scale attention heatmaps.
2. The Swin Transformer embeddings are passed to SigLIP, which generates text queries describing the identified anomaly.

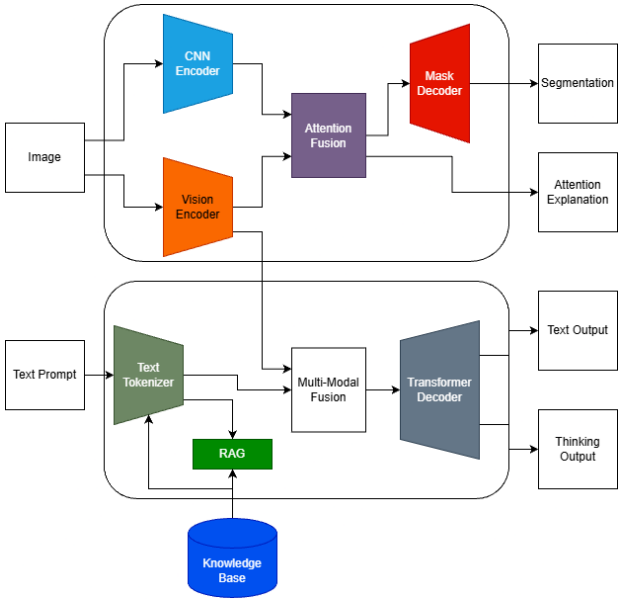

Figure 11.1: Dual Encoder + LLM

3. The queries are passed to the RAG system, which retrieves relevant medical context from the knowledge base.

4. The retrieved context and query form a system prompt for the LLM, which generates a textual explanation.

5. The segmentation mask, attention heatmaps, and textual explanation are presented to the medical professional as a unified output.

Future work should focus on implementing and validating this pipeline, constructing a domain-specific medical knowledge base for the RAG system, and evaluating the clinical utility of the generated explanations through user studies with medical professionals. Figure 11.1 depicts this architecture.

11.2 Echocardiogram Segmentation and Ejection Fraction Computation

An extension of the framework is its application to echocardiogram analysis for cardiac segmentation and ejection fraction (EF) computation. [1] Ejection fraction quantifies the percentage of blood pumped out of the left ventricle with each contraction. Ejection fraction is a key metric for diagnosing cardiovascular conditions including heart failure and cardiomyopathy. Values outside the range of 50-70% indicate significant cardiac abnormality and pathological risk.

11.2.1 Dataset

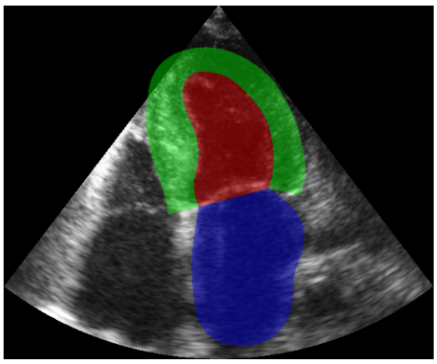

Figure 11.2: CAMUS Sample

We began a preliminary experiment on the CAMUS dataset [24], a fully annotated dataset of echocardiograms from 500 patients. Figure 11.2 displays a sample. CAMUS contains two standard views: two-chamber (2CH) and four-chamber (4CH) views. For each patient and view, two key frames are annotated: the end-diastolic (ED) frame, where the left ventricle is at maximum volume, and the end systolic (ES) frame, where the ventricle is at its smallest volume. Additionally, the dataset consists of approximately 12 frames capturing the full cycle between ED and ES for each patient. Overall, the dataset

[1] This is research that we began over the past year, in conjunction with Prof. Bentahar and his group at Concordia University

contains over 6000 frames across all patients, views, and sequences, allowing for a large dataset to train our segmentation model.

Images were preprocessed by resizing to 224×224 pixels to meet the input requirements of the Vision Transformer encoder. Pixel values were normalized from $[0, 255]$ to $[0, 1]$, and random augmentations including random blurring and scaling were applied to improve generalization and prevent overfitting.

11.2.2 The Preliminary ViT Architecture

When first imagined, the model employed a Vision Transformer encoder using a pretrained backbone with a patch size of 16×16. We utilized a lightweight decoder that projects the ViT embedding dimensions from 768 channels to 256 channels. Moreover, the projection is followed by a progressive upsampling that mimics the U-Net decoder. We restore the input resolution by progressively upsampling the feature maps back to the original 224×224 input resolution through transposed convolutional layers. The decoder was originally designed to restore the spatial resolution; however, the artifacts within each 16×16 patch were embedded into the representations prior to decoding the segmentation mask. As a result, the progressive upsampling amplified the artifacts instead of resolving them. This was the design proposed to the cardio application group and my task was to imagine which encodings might assist them the most.

11.2.3 Challenges and Findings

The experiments revealed significant challenges that ultimately prevented accurate segmentation of the left ventricle segmentation for our ViT based segmentation model.

The first challenge resulted from the inherent architecture of the patch-based nature of the ViT encoder. The model produced segmentations with a blocky, cuboid artifact pattern. The pattern was attributed to the 16×16 patch-size that Vision Transformers use; predictions were made at a patch level (see Figure 11.3) rather than a pixel wise. This resulted in jagged segmentation boundaries that are unsuitable for precise left ventricle delineation, essential for computing ejection fraction accurately.

A second challenge was the inherent poor image quality of the CAMUS dataset. The echo images contained low contrast, significant noise, and blurring. Therefore, the model's reasoning in distinguishing between adjacent cardiac chambers was distorted. As a result,

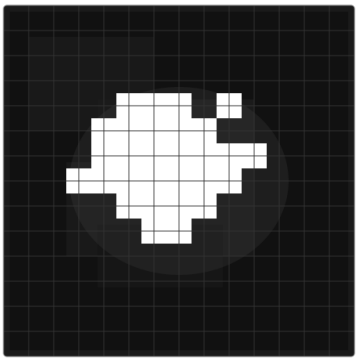

Figure 11.3: Example of patch-level segmentation artifacts produced by the ViT encoder. Each white square represents a 16×16 patch predicted as left ventricle, resulting in blocky boundaries unsuitable for precise cardiac structure delineation.

the model hallucinated by misidentifying the left atrium and left ventricle, producing fused segmentations or segmenting entire chambers incorrectly.

The third challenge emerged from training on both the 2CH and 4CH views simultaneously. The original intent was to produce a versatile model capable of handling both views. However, the anatomical differences and spatial locations between the two views introduced increased variability. Furthermore, the additional variability increased the difficulty of consistent left ventricle localization, adding noise to the learning process.

11.2.4 LLM Prototype and Textual Explanation

To explore the textual explanation component of the proposed multimodal framework, we went further. Here is our preliminary design. We demonstrate a prototype using DeepSeek-R1 [9] as the primary LLM model. An arbitrary yet realistic ejection fraction value and contextual information about the echocardiogram is prompted to the network. The system subsequently generates both a chain of thought reasoning as an explanation and a clinical report for the medical professional.

The system prompt used to the guide the LLM is designed with negative prompting to separate the role of the AI system from a medical professional and prevent the model from overstepping its boundaries. The prompt is as follows:

> *You are an AI assistant that helps summarize medical imaging findings for qualified healthcare professionals. You are not a medical expert and you are not*

making any clinical decisions. Your role is to present the findings from the segmentation model in a clear and structured way to assist the medical professional in their assessment. You do not diagnose, prescribe, or recommend treatment. All clinical decisions must be made by the qualified medical professional reviewing this report.

The use of negative prompting in the system is critical for clinical deployment, as it prevents the model from generating overconfident or misleading clinical recommendations. Research demonstrates that confining the role of the AI system in clinical decision support can improve the medical professional trust and reduce the reliance on AI [32].

The demonstration was to produce two outputs. The first is chain of thought reasoning, in which the model reasons step by step based on clinical implications of the provided EF value, considering the appropriate normal range. The second is a clinical report summarizing the findings for the medical professional, including the EF value, implications, view considerations, and recommendations.

A structured report of the LLM (DeepSeek-R1) output for a patient with an EF of 40% is as follows:

> Assessment of Ejection Fraction (EF):: The patient's ejection fraction is 40%, which measures the percentage of blood pumped out by the left ventricle with each contraction.
>
> Normal Range: A normal EF typically falls between 50-70%. Therefore, an EF of 40% is below this range. Implications: An EF below the normal range may indicate conditions such as heart failure (HF with reduced ejection fraction), cardiomyopathy, or other structural heart issues. This suggests that the left ventricle is not contracting effectively to pump out the usual amount of blood during each heartbeat.
>
> View Considerations: The 4CH view provides a comprehensive look at all four chambers of the heart, which should give a good overall assessment. However, the specific phases (End Systole and End Diastole) are used in calculating the EF, and they do not directly influence the measurement beyond providing necessary data points for calculation.
>
> Next Steps: While an EF of 40% is concerning and suggests potential heart function issues, further evaluation by a healthcare professional is required to determine the exact cause and appropriate management. This may include reviewing the echocardiogram images, considering other physiological factors

that could affect the measurement, or assessing for additional clinical symptoms or conditions. In conclusion, an EF of 40% indicates suboptimal heart function and warrants further medical evaluation to address potential underlying issues.

The thinking output, which is the textual explanation of how the LLM decided its report is as follows:

Thinking... Alright, so I'm trying to figure out how to assess this patient's ejection fraction based on their echocardiogram results. They mentioned that the EF is 40%, and the images are from both End Systole and End Diastole in a 4CH view.

First, I need to recall what an ejection fraction (EF) represents. From what I remember, it's a measure of how much blood the left ventricle pumps out with each contraction. So, if the EF is 40%, that means the left ventricle is pumping out 40% of its volume during each heartbeat. Now, what's considered a normal ejection fraction? Earlier in our conversation, I noted that a normal range is typically between 50-70%. That seems to be a standard reference point. So, an EF of 40% falls below this range.

I should think about what could cause a lower than normal ejection fraction. Conditions like heart failure especially HF with reduced ejection fraction), cardiomyopathy, or other structural heart issues might lead to a decreased EF. But I'm not entirely sure how the 4CH view affects this measurement. The patient mentioned that the images are from both End Systole and End Diastole phases in the 4CH view. The 4CH view provides a comprehensive look at all four chambers of the heart, which should give a good overall assessment. But does the phase (end systole or end diastole) influence how EF is calculated? I think the calculation typically uses measurements from these specific points to determine the volume ejected.

So, if the EF is 40%, it's below normal. This could indicate that the left ventricle isn't contracting as effectively as it should be, reducing the amount of blood pumped out with each beat. But without seeing the actual images or additional details, I can't pinpoint the exact cause. It might require further evaluation by a healthcare professional.

I also need to consider if there are any factors that could falsely lower the EF reading. Maybe image quality issues in the echocardiogram? Or perhaps other physiological conditions affecting the measurement. But generally, an EF of 40% is concerning and suggests some form of ventricular dysfunction.

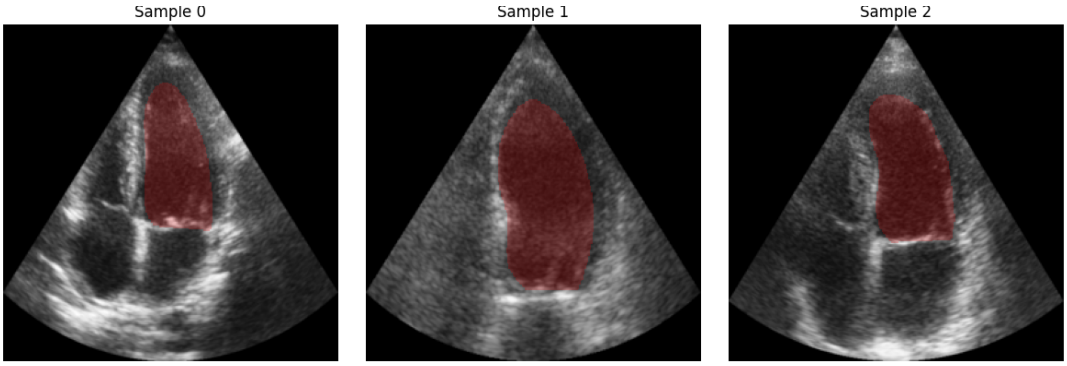

Figure 11.4: Example of Left Ventricle Segmentation

> In summary, based on the EF of 40%, it's below the normal range, indicating potential heart function issues that need to be addressed by a medical professional. ...done thinking.

The outputs demonstrate the feasibility of generating textual reports from a given segmentation for clinicians and supporting the report with a chain of thought textual explanation. However, a key limitation identified during the prototype is that the LLM does not account for patient-specific context such as age and gender, which are clinically relevant factors that can influence the interpretation of EF measurements. This represents an important direction for future refinement of the system prompt and the RAG knowledge retrieval pipeline.

11.2.5 Future Directions

The findings from the preliminary experiment directly motivated the application of the dual encoder architecture from Part Two for this task. In fact, the ViT model served as a predecessor for the dual encoder architecture. The blocky patch artifacts produced by the standalone ViT encoder are due to the absence of the CNN encoder to capture local spatial details at the pixel level. The dual encoder architecture utilized ResNet-34 CNN encoder for the local feature extraction to fix the artifacts and refine edge detections instead of blocks. Unlike the standard ViT, the Swin Transformer produces hierarchical multi-scale feature maps, thus skip connections to can be naturally extracted at each encoder stage and fused with the corresponding CNN features. Furthermore, the ablation study presented in Part Two confirmed that the Swin Transformer consistently outperforms ViT variants

across all evaluated configurations. Beyond the segmentation, the embeddings the Swin Transformer encoder produces can be leveraged alongside a language model to effectively form the same vision language model.

With respect to the textual explanation explanation component, the system prompt represents area for refinement. Patient-specific context such as age, gender, and medical history into the prompt would enable the LLM to generate more personalized and clinically relevant explanations and reports. Furthermore, the RAG knowledge base may be implemented to retreive patient-specific context and further grounding the textual explanations in medical context. Additionally, evaluating and fine tuning multiple LLMs including models such as GPT [1], Gemma [43], and LLama [44] would provide better insights on which model may be best suited for clinical text generation for this application.

We propose evaluating the dual encoder framework on the CAMUS dataset [24]. The framework would segment the left ventricle across the cardiac cycle to compute EF, while providing visual attention heatmaps. Furthermore, we propose to extend the dual encoder framework by extracting embeddings from the Swin Transformer encoder to produce textual explanations, effectively forming a Vision-Language Model (VLM) that describes the segmentation findings and their clinical implications.

Echocardiogram segmentation presents distinct challenges compared to polyp segmentation. The left ventricle is a singular contiguous structure with a consistent shape, where local CNN features may be sufficient for boundary delineation in individual frames. However, understanding ventricular motion across the cardiac cycle requires integrating information over a larger spatial and temporal context, motivating the global modeling of the Swin Transformer encoder. This application provides an interesting experiment for evaluating how the dual encoder balances local and global information for a distinct segmentation task.

11.2.6 Breast Cancer Segmentation

Another extension is the application of the framework to breast cancer tumor segmentation. Breast cancer is one of the most prevalent cancers worldwide, and accurate tumor identification is essential for treatment planning. We propose evaluating the framework on the MAMA-MIA dataset [14], which contains breast cancer DCE-MRI scans with expert tumor segmentations. Beyond segmentation, the multimodal framework would provide visual and textual explanations describing potential lesion subtypes, including luminal A, luminal B, HER2-enriched, and triple-negative, each requiring a distinct treatment response. We performed an initial investigation of this application and acquired additional

important insights. The dataset in question was large, and compute power was insufficient to carry out all the intended processing. This suggests that steps forward might consider either adding more compute power, making use of synthetic datasets, or only operating on a subset of the existing real world dataset.

11.2.7 Attention Map Aggregation

The dual encoder architecture generates individual attention maps at each of the four fusion levels (see Figure 7.1). A useful direction for future work is aggregating these multi-scale maps into a single composite explanation that is easier to interpret clinically. Candidate strategies include simple averaging across decoder levels, weighted combination that emphasizes certain levels based on the task, and a learned aggregation mechanism that optimally combines maps for both performance and explanation quality. User studies with medical professionals would help determine which strategy best supports clinical decision making.

11.2.8 Recommendations for Next Generation of Medical Datasets

Many current medical imaging datasets are limited in size and image quality, a significant challenge for training deep learning models effectively. Additionally, the majority of available datasets are designed for classification tasks, with fewer providing pixel-wise segmentation masks for training segmentation models. This limits the progress of research in medical image segmentation and explainability.

The next generation of medical datasets should address these limitations by providing not only images and segmentation masks, but also rich annotations in the form of textual descriptions, clinical notes, and structured metadata such as patient demographics and imaging conditions. Future datasets that pair visual data with textual annotations would enable researchers to build stronger multimodal models that can generate both visual and textual explanations. Additionally, such datasets would be valuable for advancing the RAG and LLM based textual explanation pipeline proposed in this thesis. Rich knowledge base grounded in annotated medical data would significantly improve the quality and clinical relevance of the generated explanations.

11.2.9 Being Attuned to Medical Experts

With respect to the needs of medical professionals for explainable AI, some related work suggests important avenues to explore. We have already mentioned the step of designing user studies involving medical professionals. Panigutti et al. [32] with their engagement of medical experts show that explanations in clinical decision support can increase advice-taking. Included are efforts to obtain qualitative feedback (comments provided by clinicians to accompany any quantitative evaluation) as an excellent strategy for gaining important insights. The authors also reveal some areas where there may be hesitations: clinicians may request explanations, yet report cognitive overload when interfaces become too dense. This is a point also brought to light by Zhao et al. [52] who show that graphical perception factors, such as visual encoding choices, substantially shape how people interpret saliency-based explanations. With respect to the heatmap component of our work, the point is that even though our attention-based architecture will offer improved explainability, clinical utility will still depend on visualization design as we move forward with new directions. All the same we are encouraged by recent controlled evidence suggesting that explainability can improve task outcomes in expert human-AI collaboration. Senoner et al. [39] report measurable gains in expert performance when heatmaps accompany AI outputs compared to black-box assistance [39].

Concerns about over-reliance on AI outputs and the replacement of clinical judgment also remain salient [32]. This reinforces the importance of the negative prompting strategy discussed in this thesis, where the system prompt explicitly constrains the role of the AI system and reminds the clinician that all final decisions rest with them.

Another thread for future work is to study existing datasets to see whether the way they are coded may pose challenges to effective segmentation. In preliminary experiments of ours on the CAMUS dataset for left ventricular segmentation [24] and with a simpler ViT architecture, the greyscale nature of the echocardiogram images made it challenging to differentiate between the left ventricle, myocardium and surrounding cardiac structures. Future work can explore whether certain datasets will be inherently more difficult to use, perhaps leading to efforts to expand the data collection process towards better outcomes for our medical professionals.

11.3 Conclusion

This thesis presented two frameworks for explainable medical image segmentation, motivated by the need for AI systems that are accurate, interpretable, and trustworthy in

clinical practice.

Part One introduced an attention-based CNN architecture that achieves competitive segmentation performance while generating model-specific visual explanations through spatial attention gates. The key finding is that model-specific explanations generated during the forward pass are more faithful to the model's reasoning than post-hoc methods, and that accuracy and explainability can be achieved simultaneously.

Part Two advanced this by introducing a dual encoder architecture combining a ResNet-34 CNN encoder and a Swin Transformer encoder, fused through a learned directional attention-gated mechanism at multiple scales. The dual encoder achieves state-of-the-art performance on Kvasir-SEG, with a Dice score of 0.907, outperforming all single encoder baselines. The Swin-guided attention configuration consistently outperforms all other configurations, demonstrating that combining local CNN features with global transformer context through directional attention-gated fusion is both an effective segmentation strategy and a principled approach to explainability.

Part Three outlined a pathway for extending the framework towards a fully multimodal explainability system, integrating textual explanations through SigLIP, RAG, and an LLM alongside the visual heatmaps. Future directions include echocardiogram segmentation and ejection fraction computation, breast cancer segmentation, attention map aggregation, and user studies with medical professionals.

We believe that accuracy and interpretability are not competing objectives. The results in this thesis demonstrate that model-specific explainability is a natural consequence of thoughtful architectural design. We hope the frameworks and future directions outlined here contribute to the responsible integration of AI into clinical workflows and ultimately lead to better outcomes for patients.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Saleh Albelwi and Ausif Mahmood. A framework for designing the architectures of deep convolutional neural networks. *Entropy*, 19(6):242, 2017.

[3] Jorge Bernal, Francisco Javier Sánchez, Glòria Fernández-Esparrach, Debora Gil, Cristina Rodríguez de Miguel, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society*, 43:99–111, 2015.

[4] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

[5] Dong Bo, Wang Wenhai, Fan Deng-Ping, Li Jinpeng, Fu Huazhu, and Shao Ling. Polyp-pvt: Polyp segmentation with pyramidvision transformers. *CAAI AIR*, 2023.

[6] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018.

[7] Jieneng Chen, Jieru Mei, Xianhang Li, Yongyi Lu, Qihang Yu, Qingyue Wei, Xiangde Luo, Yutong Xie, Ehsan Adeli, Yan Wang, Matthew P. Lungren, Shaoting Zhang, Lei Xing, Le Lu, Alan Yuille, and Yuyin Zhou. Transunet: Rethinking the u-net

architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis*, 97:103280, 2024.

[8] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[9] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Jun-Mei Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiaoling Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bing-Li Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, and Chong Ruan et Al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645:633 – 638, 2025.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[11] Foivos I Diakogiannis, François Waldner, Peter Caccetta, and Chen Wu. Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162:94–114, 2020.

[12] Alexey Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[13] Razvan-Gabriel Dumitru, Darius Peteleaza, and Catalin Craciun. Using duck-net for polyp image segmentation. *Scientific reports*, 13(1):9803, 2023.

[14] Lidia Garrucho, Smriti Joshi, Kaisar Kushibar, Richard Osuala, Maciej Bobowicz, Xavier Bargall'o, Paulius Jarusevicius, Kai Geißler, Raphael Schäfer, Muhammad Alberb, and Tony Xu et Al. The mama-mia challenge: Advancing generalizability and fairness in breast mri tumor segmentation and treatment response prediction. *ArXiv*, abs/2603.01250, 2026.

[15] Iryna Hartsock and Ghulam Rasool. Vision-language models for medical report generation and visual question answering: A review. *Frontiers in artificial intelligence*, 7:1430984, 2024.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[17] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[18] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas De Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *International conference on multimedia modeling*, pages 451–462. Springer, 2019.

[19] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Dag Johansen, Thomas De Lange, Pål Halvorsen, and Håvard D Johansen. Resunet++: An advanced architecture for medical image segmentation. In *2019 IEEE international symposium on multimedia (ISM)*, pages 225–2255. IEEE, 2019.

[20] Mahmoud Khalaf, Robin Cohen, Paulo Alencar, Jamal Bentahar, and Saidul Islam. A little attention: Model-specific explainable medical imaging. In *Proceedings of the IJCAI 2025 Workshop on Large Language Models and Generative AI for Health Informatics*, pages 1–9, 2025.

[21] Mahmoud Khalaf, Robin Cohen, Saidul Islam, and Jamal Bentahar. Explainable medical image segmentation via attention-gated fusion of ViT and U-Nets. In *Proceedings of the 39th Canadian Conference on Artificial Intelligence*, volume 318 of *Proceedings of Machine Learning Research*. PMLR, 2026.

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[24] Sarah Leclerc, Erik Smistad, Joao Pedrosa, Andreas Østvik, Frederic Cervenansky, Florian Espinosa, Torvald Espeland, Erik Andreas Rye Berg, Pierre-Marc Jodoin, Thomas Grenier, et al. Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. *IEEE transactions on medical imaging*, 38(9):2198–2210, 2019.

[25] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

[26] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[27] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[28] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[29] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature communications*, 15(1):654, 2024.

[30] Subhankar Maity and Manob Jyoti Saikia. Large language models in healthcare and medical applications: a review. *Bioengineering*, 12(6):631, 2025.

[31] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas. In *Medical Imaging with Deep Learning*, 2018.

[32] C. Panigutti, A. Beretta, F. Giannotti, and D. Pedreschi. Understanding the impact of explanations on advice-taking: A user study for ai-based clinical decision support systems. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2022.

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[34] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD*

international conference on knowledge discovery and data mining, pages 1135–1144, 2016.

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[36] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[37] Stuart J. Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson, 4 edition, 2020.

[38] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[39] J. Senoner et al. Explainable ai improves task performance in human–ai collaboration. *Scientific Reports*, 14:31150, 2024.

[40] RL Siegel, KD Miller, HE Fuchs, and A Jemal. Cancer statistics, 2022. *CA: a Cancer Journal for Clinicians*, 72(1):7–33, 2022.

[41] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015.

[42] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[43] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

[44] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal

Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[46] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[47] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, F. Xia, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

[48] R. Yang, Y. Ning, E. Keppo, et al. Retrieval-augmented generation for generative artificial intelligence in health care. *npj Health Systems*, 2:2, 2025.

[49] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023.

[50] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5625–5644, 2024.

[51] Yundong Zhang, Huiye Liu, and Qiang Hu. Transfuse: Fusing transformers and cnns for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 14–24. Springer, 2021.

[52] Y. Zhao, M. Li, and M. Berger. Graphical perception of saliency-based model explanations. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2023.